

NETWORKED TRUST: COMPUTATIONAL
UNDERSTANDING OF INTERPERSONAL TRUST
ONLINE

A Dissertation

Presented to the Faculty of the Graduate School
of Cornell University

in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy

by

Xiao Ma

August 2019

© 2019 Xiao Ma

ALL RIGHTS RESERVED

NETWORKED TRUST: COMPUTATIONAL UNDERSTANDING OF
INTERPERSONAL TRUST ONLINE

Xiao Ma, Ph.D.

Cornell University 2019

My doctoral research develops a **deeper understanding of interpersonal trust online through computational methods**, in the context of online exchange platforms including peer-to-peer marketplaces, sharing economy platforms, and social networks. Through analyzing images in product listings on eBay and LetGo.com, language in profiles on Airbnb, and networks in social groups on Facebook, I show how different algorithms help understand and predict interpersonal trust in each context. Findings reveal patterns of interpersonal trust. For example, high-quality images are perceived as more trustworthy than stock imagery; language of promises lead to higher perceived trustworthiness through conventional signaling; and smaller, denser, and more private social groups are trusted more. These findings inform the design of online exchange platforms. The algorithms predicting trust could also be used for better ranking and recommendation to “engineer” interpersonal trust. Going forward, I propose a lens of “**networked trust**” to view interpersonal trust online, which has three focuses: (1) *cues* in Computer-Mediated Communication; (2) embeddedness in social *networks*; and (3) increasing mediation by *algorithms*. The networked trust framework can be used to frame future trust research in other contexts, such as **misinformation**. Finally, two research agenda were charted by this dissertation — **AI-Mediated Communication** and **AI-Mediated Exchange Theory**, which future work can develop on.

BIOGRAPHICAL SKETCH

Xiao Ma is a PhD candidate in Information Science and a member of the Social Technologies Lab at Cornell Tech, Cornell University. She started the PhD in September 2014 at Cornell Tech’s temporary campus at the Google NYC building in Chelsea before it moved to Roosevelt Island in July 2017. She is the first PhD to graduate fully contained at the new campus of Cornell Tech since its establishment in 2012.

During her PhD, Xiao has worked as a User Experience (UX) researcher at Airbnb and Facebook, and as a data scientist/engineer at the Core Data Science team at Facebook. Different roles have given her different perspectives on the design, implementation, and analysis of real-world socio-technical systems. She also holds a Bachelor of Science in Microelectronics from Peking University.

To my parents, for giving me the courage and freedom to see a world of infinite possibilities.

ACKNOWLEDGEMENTS

The past five years at Cornell Tech have been such an incredible journey. I had a front-row seat to witness the founding and the growth of Cornell's campus in New York City, Cornell Tech. I would like to thank everyone who made Cornell Tech possible. It was an honor to see the amount of work that goes into creating something new from scratch, and experience the change as we grow. Entrepreneurship is a large part of my identity in addition to research, and I hope to get a chance to build something from scratch of my own one day.

I believe the best part of the PhD is the people, and I have so many to thank.

First and foremost, my deepest gratitude goes to my advisor, Mor Naaman, who not only taught me how to do good research, but also how to be a better human being. Thank you for always seeing my potential and encouraging me to achieve it. One of the most important lessons I learned from Mor is that the questions matter equally if not more than the answers. I love the process of brainstorming and discussing with Mor to formulate a question, articulate it, then finally solve it. I also thank Mor for sharing his travel stories and introducing me to more music. I probably didn't travel as much as Mor did during his PhD, but I definitely have been to more art exhibits.

I am also so fortunate to have a very interdisciplinary and supportive committee, including Jeff Hancock, Karen Levy, and Serge Belongie. Thank you for teaching me on different fields, including communication, sociology, and computer science. I'd also like to thank Natalie Bazarova, Clarence Lee, Louise Barkhuus and Raz Schwartz for providing important feedback on my early work during the PhD.

Outside of Cornell, I have also learned a lot from having support from other institutions, including Stanford, Facebook, Airbnb, and Women in Technology

and Entrepreneurship in New York (WiTNY).

I visited the Stanford Institute for Research in the Social Sciences in my second year of PhD. Thank you to Paolo Parigi and Bruno Abrahao for hosting me, and Michael Bernstein for having me spend time with the Stanford HCI group. I made good friends there including Ali Alkhatib, Niloufar Salehi, Ranjay Krishna, and many others. I also reconnected with Scott Cheng, who was doing his Master's in HCI at Stanford at the time and later became my partner when he moved to New York. Scott and I have very different personalities. His calmness is always there for me whenever I experience anxieties from research.

Facebook and Airbnb opened the door for industry for me. I am grateful for having experienced the unique cultures at both places. Thank you to Judd Antin, Janna Bray, Yoni Karpfen, and many others at Airbnb for showing me the ropes of industry research. Thank you to Jenna Lee, Tracy Mehlman, and Affonso Reis for helping me think more deeply about career trajectory. Finally, thank you to the Facebook Core Data Science team, especially Justin Cheng, Lada Adamic, Shankar Iyer, Alex Dow, Moira Burke, Bogdan State, Carlos Diuk, for one of my favorite summers in California. I have always admired Justin's work and it was an honor to be able to work together. One of my favorite memories during that summer is the Computational Social Science group's research writing retreat Lada organized, when we all discussed research ideas around bonfire in the mountains of Santa Cruz.

I was also so fortunate to be a Women in Technology and Entrepreneurship in New York (WiTNY) fellow during my PhD, advised by Judy Spitz. Through working with Judy, I developed more leadership skills and made efforts to contribute back to the community. Judy is charismatic, resourceful. She is an amazing leader, mentor, and role-model.

I also want to thank all my lab mates and friends I made during graduate school, including Emily Sun, Ross McLachlan, Nir Grinberg, Maurice Jakesch, Kimberly Wilber, Hani Altwaijry, Andreas Veit, Neta Tamir, Justine Zhang, and many others. I thank Amy X. Zhang at MIT, who convinced me to come to Cornell Tech to work with Mor, and who has always been a role-model and inspiration.

Finally, I would like to thank my parents, Jianping Ma and Yan Liang. My father is an academic himself and I grew up on the campus of his university, spending more time in the libraries and classrooms than on the playground. He inspired me to be intellectually curious, and to have broad interests (he is a physicist with interest in calligraphy, poetry and literature). My mother studied computer science, and worked as a software engineer in the 1970s in China. She raised me to be independent, strong, and taught me to write my first program. My parents instilled in me by example the best ethics, in life, and about work. They also gave me the freedom to travel, to see the world, and most importantly, to explore all the possibilities and make my own decisions — they even gave me a name that means freedom in Chinese. The freedom is especially valuable coming from Chinese parents of their generation because of the single child policy. Confucius says, "Don't travel far away while your parents are still alive. If you do, make it worthwhile." I hope I have made the PhD journey worthwhile, and I hope my parents are proud and may they forgive me for all the time I am absent away from home.

TABLE OF CONTENTS

| | |
|---|-----------|
| Biographical Sketch | iii |
| Dedication | iv |
| Acknowledgements | v |
| Table of Contents | viii |
| List of Tables | x |
| List of Figures | xi |
| 1 Introduction | 1 |
| 1.1 Outline | 10 |
| 1.2 Summary of Contribution | 12 |
| 2 Background | 15 |
| 2.1 Theoretical Foundations of Trust | 16 |
| 2.2 Measuring Trust | 21 |
| 2.2.1 Determinants and Outcomes of Trust — A Multilevel Perspective | 23 |
| 2.2.2 Methodological Note on Measuring Trust | 32 |
| 2.3 Online Trust and Networked Trust | 37 |
| 2.3.1 Online Trust | 37 |
| 2.3.2 Networked Trust | 40 |
| 2.4 Bias and Distrust | 43 |
| 2.5 Summary | 46 |
| 3 Images of Trust: Understanding Image Quality and Trust in Peer-to-Peer Marketplaces | 48 |
| 3.1 Introduction | 48 |
| 3.2 Related Work | 52 |
| 3.3 Datasets | 54 |
| 3.3.1 LetGo.com | 55 |
| 3.3.2 eBay | 55 |
| 3.4 Annotating Image Quality | 56 |
| 3.5 Modeling Image Quality | 58 |
| 3.5.1 What Makes a Good Product Photo? | 60 |
| 3.6 Marketplace Outcomes | 67 |
| 3.6.1 Image Quality and Sales | 68 |
| 3.6.2 Perceived Trustworthiness | 70 |
| 3.7 Discussion and Conclusion | 73 |
| 4 Language of Trust: Self-Presentation and Perceived Trustworthiness of Airbnb Host Profiles | 77 |
| 4.1 Introduction | 77 |
| 4.2 Related Work | 80 |

| | | |
|----------|---|------------|
| 4.3 | Step 1: Self-Presentation and Perceived Trustworthiness of Airbnb Host Profiles | 82 |
| 4.3.1 | Study 1 — How do Hosts Self-Disclose? | 85 |
| 4.3.2 | Study 2 — Self-Disclosure and Perceived Trustworthiness | 96 |
| 4.3.3 | Study 3 — From Perception to Choice | 106 |
| 4.4 | Step 2: A Computational Approach to Perceived Trustworthiness of Airbnb Host Profiles | 111 |
| 4.4.1 | Method and Dataset | 113 |
| 4.4.2 | Predicting Perceived Trustworthiness | 115 |
| 4.4.3 | Factors Contributing to Perceived Trustworthiness | 120 |
| 4.5 | Discussion | 120 |
| 4.5.1 | Design Implications | 125 |
| 4.5.2 | Limitations | 125 |
| 4.6 | Conclusion and Extensions | 127 |
| 5 | Networks of Trust: When Do People Trust Their Social Groups? | 131 |
| 5.1 | Introduction | 131 |
| 5.2 | Related Work | 135 |
| 5.2.1 | Determinants of Trust in Groups | 135 |
| 5.3 | Methods | 137 |
| 5.3.1 | Sampling | 138 |
| 5.3.2 | Survey Design | 139 |
| 5.3.3 | Data and Statistical Approaches | 140 |
| 5.4 | Results | 142 |
| 5.4.1 | Individual Differences and Trust | 143 |
| 5.4.2 | Group Differences and Trust | 144 |
| 5.4.3 | Predicting Trust in Groups | 153 |
| 5.4.4 | Group Outcomes | 154 |
| 5.5 | Discussion | 156 |
| 5.5.1 | Design Implications | 158 |
| 5.5.2 | Limitations | 159 |
| 5.6 | Conclusion | 160 |
| 6 | Discussion and Future Work | 162 |
| 6.1 | Algorithms of Trust and Networked Trust | 162 |
| 6.2 | Future Research Agenda | 164 |
| 6.2.1 | Networked Trust and Misinformation | 164 |
| 6.2.2 | AI-Mediated Communication (AI-MC) | 168 |
| 6.2.3 | AI-Mediated Exchange Theory (AI-MET) | 170 |
| 6.3 | Conclusion | 174 |
| | Bibliography | 177 |

LIST OF TABLES

| | | |
|-----|--|-----|
| 3.1 | Image feature definitions and example images | 63 |
| 3.2 | Ordered logistic regression coefficients predicting image quality labels | 66 |
| 3.3 | Image quality predicted by our models is positively associated with higher likelihood that an item is sold (1.17x more for shoes, and 1.25x more for handbags). | 69 |
| 3.4 | Marketplace perceived trustworthiness scale | 74 |
| 4.1 | Topics of self-disclosure in Airbnb host profiles. | 87 |
| 4.2 | Six-item perceived trustworthiness scale. | 98 |
| 4.3 | Performance of sentence category classification. | 118 |
| 4.4 | Model performance (accuracy) summary by different length batch. Random baseline accuracy is 50%. Models compare profiles of similar lengths to predict relative perceived trustworthiness based on word count. | 119 |
| 4.5 | Factors contributing to higher and lower perceived trustworthiness by different length batch. LIWC categories “we” and “social” are important positive indicators, potentially through social warranting. The LIWC “sexual” category contained mostly the word “love”. | 130 |
| 5.1 | Trust in groups survey items. Participants reported the degree to which they agreed or disagreed to each of the survey items on a five-point Likert scale. | 141 |
| 5.2 | Descriptive summary of survey measures, including general attitudes and trust in groups. Sparklines represent the histogram of each measure. (N=6,383) | 142 |
| 5.3 | Baseline model predicting trust in groups using demographics, disposition to trust, risk attitude, in-group loyalty, and social support. (N=6,323 after removing missing age and gender observations) | 143 |
| 5.4 | Five sets of group-level features used for predicting trust in groups. | 145 |

LIST OF FIGURES

| | | |
|-----|--|-----|
| 2.1 | Direct and two forms of generalized (indirect) exchange illustrated by Bearman 1997. Each character represents a party involved in the exchange. In direct social exchange, there is direct transfer of value or goods between A and B. In generalized exchange, multiple parties pool their resources together to produce greater value, which is then further distributed. | 19 |
| 2.2 | Differences among trust, trustworthiness and perceived trustworthiness. | 22 |
| 2.3 | Multilevel perspective of trust. | 24 |
| 2.4 | The integrative organizational trust model proposed by Mayer et al. 1995 | 31 |
| 3.1 | We study the interplay between image quality, marketplace outcomes, and user trust in peer-to-peer online marketplaces. Here, we show how image quality (as measured by a deep-learned CNN model) correlates with user trust. User studies in Sec. 3.6.2 show that high quality images selected by our model out-performs stock-imagery in eliciting user trust. | 51 |
| 3.2 | Samples of lowest-rated and highest-rated images from shoes and handbags groundtruth. | 52 |
| 3.3 | Left: standardized score distributions for filtered images on shoes and handbags categories. Right: final ground truth labels, showing a fairly even distribution. | 57 |
| 3.4 | Hypothetical marketplace mock-ups used for our user experiment. From left to right, showing images with high quality score, low quality score, and stock imagery. | 70 |
| 4.1 | Probability of disclosure per topic | 94 |
| 4.2 | Average number of sentences per topic | 94 |
| 4.3 | Self-disclosure trends by topic and host type. The error bars represent one standard error. | 94 |
| 4.4 | Perceived trustworthiness increases with profile length (x -axis on log scale). | 101 |
| 4.5 | Perceived trustworthiness score distributions for profiles with different number of topics. | 103 |
| 4.6 | Comparison for different strategies, organized by the number of topics mentioned in the profile. The dotted lines indicate the bottom and top quartiles in each topic-count group. | 104 |
| 4.7 | Observed likelihood that the host with high trustworthiness score is preferred and the probability predicted by Bradley-Terry model, by profile length. The error bars represent one standard error. . . | 110 |

| | | |
|-----|--|-----|
| 5.1 | The relationship between trust in groups and group size, for each dimension (panels), across groups of different privacy types (line style) and individuals with different propensity to trust (line color). Dunbar’s number (150) is marked by a vertical red dotted line.) | 147 |
| 5.2 | People have the highest trust in friends and family groups, and lowest in interest- and location-based groups. | 149 |
| 5.3 | Groups differ in network density, participant degree centrality, and how a participant’s friends are linked to each other. Each node represents a group member. Each edge represents a friendship between two members. The survey participant is colored in red, and group admins are colored in yellow. | 152 |
| 5.4 | For each feature set, we calculated the average feature importance (measured by relative percent increase in MSE when a feature is removed) in predicting trust in groups. Network structure was the most important, followed by an individual’s general attitudes towards others. | 153 |
| 5.5 | Groups with higher trust ratings are less likely to increase in size (left), more likely for the survey participant to form new connections in them (right), and had no effect on the likelihood on forming friendships among group members other than the rater (center). | 155 |

CHAPTER 1

INTRODUCTION

Trust, in the broadest sense of confidence in one's expectations, is a basic fact of social life.

Niklas Luhmann
in *Trust and Power* (1979)

Our world builds on the foundation of **social exchange**. When we say social exchange, we refer to both the psychological and economic aspects of social relationships [Emerson, 1976]. According to Homans, elementary forms of social behavior such as communication among group members can be viewed as an exchange [Homans, 1958]. Employment relationships can also be viewed as an exchange where the reward to the employee is the wage, and the cost is the responsibilities they have [Homans, 1958]. Viewing social relations as exchanges, especially as exchanges in social networks, is helpful in understanding phenomena in social structures, including group cohesiveness [Homans, 1958], social solidarity [Bearman, 1997], and power [Blau, 1964].

One of the most important concept that enables the functioning of social exchange systems at scale is **trust**. Because there are often **risks** involved in a potentially mutually beneficial exchange, trust is needed to overcome the risks and to facilitate the exchange. For example, commerce could potentially be beneficial for both the buyer and the seller. But the buyer must take “a leap of faith” and trust the seller to provide accurate information about the merchandise before making the decision to purchase. Without trust, no exchange could take place under risks and hardly any form of collaboration could exist. For each

individual, without trust, the everyday social life which we take for granted is simply not possible. As Niklas Luhmann wrote in opening chapter of his foundational book *Trust and Power* [Luhmann, 1979]:

“[...] a complete absence of trust would prevent him or her [a person] from even getting up in the morning. [...] Such abrupt confrontation with the complexity of the world at its most extreme is beyond human endurance.”

As trust mitigates the uncertainty and complexity in social exchange systems, trust has been frequently attributed for systems' success. A deeper understanding of trust is important for the social, economic and political outcomes of social systems — including for individuals, groups, and nations. For example, trust is considered as an important variable in the development of healthy family relationships and personalities for individuals [Rotter, 1967] Trust has also been shown to be important for the productivity and cohesiveness of organizations [Mayer et al., 1995, Nyhan, 2000, McEvily et al., 2003, Fine and Holyfield, 1996]. Finally, Fukuyama showed that trust leads to the creation of prosperity for nations [Fukuyama, 1995].

Because of the importance of trust, a robust line of research exists across disciplines on the factors that lead to trust, and the outcomes that trust contribute to, in sociology, psychology, economics, marketing, and management science. At the same time, **new questions about trust**, especially trust in socio-technical systems, emerge as a result of increasing digital mediation on social exchange relationships. In the past few decades, social exchange relationships have undergone fundamental changes. In my view, **the digitalization of social exchange took place in three waves**: (1) the digitization of the exchange of *goods*; (2) the

digitization of social *relationships*; and (3) the digitalization of the exchange of *resources* through sharing economy platforms. Each of these waves brought significant changes to social exchange structures, raising new questions about trust at the same time. The work presented in this dissertation addresses these new questions through the development of computational methods.

What are some of the new questions about trust that each wave of the digitalization of social exchange brought?

The first wave, **the digitization of the exchange of goods**, was enabled by online commerce platforms including Amazon (founded in 1994) and eBay (founded in 1995). Through these digital exchange platforms, sellers can reach customers widely regardless of location, and customers in turn have much wider selections to choose from compared to their local stores. However, one core risk emerged — if the buyers pay first, it is possible that the sellers would not send the goods (and disappear from the Internet and reappear under a different name), or send defect goods that do not meet the expectations of the buyers; however, if the sellers send the goods first, the buyers could refuse to pay. In other words, people have to trust others with their *financial resources* for exchanges to take place on these online commerce platforms. Multiple mechanisms emerged to mitigate the risks and to increase trust, including the development of third-party online payment systems that acted as an assurance (e.g., PayPal, founded in 1998). Reputation systems also appeared as a key mechanism to increase trust. A long line of research on reputation systems has discussed their benefits and limitations [Resnick et al., 2000, Resnick and Zeckhauser, 2002, Resnick et al., 2006, Cook et al., 2009].

Two decades later, new questions about trust on these online commerce

platforms are still emerging as platforms continue to evolve. For example, due to the wide accessibility of mobile photography, user-generated images of goods play an ever increasingly important role in establishing trust in peer-to-peer marketplaces, compared to stock imagery. New advancements in computer vision (e.g., deep learning) also present opportunities to understand how user-generated images establish trust. This dissertation presents a case study on leveraging new computational opportunities to understand and predict how user-generated images impact trust in online marketplaces.

The second wave of the digitalization of exchange was **the digitization of social relationships**, characterized by the ubiquitous adoption of social network sites (SNSs) such as Facebook (founded in 2004), Twitter (founded in 2006), LinkedIn (founded in 2002) and Instagram (founded in 2010). Although online communities have long existed since the beginning of the Internet, these SNSs enable users to articulate and make visible their social networks [Boyd and Ellison, 2007]. New questions around trust again emerged. The core risk presented by SNSs in social exchange is how personal information flows through networks. Self-disclosure on SNSs can be visible to a “networked audience” [Marwick and Boyd, 2011], including real and potential viewers through a broadcasting audience network. Therefore, there are a lot of risks when people post or share content online. People need to trust others with their *personal information*, including information about their identities, biometric data, social ties, experiences, feelings, emotions, etc.

Important opportunities are also brought by SNSs to better understand interpersonal trust, especially around how tie strength and network structure contribute to interpersonal outcomes. For example, computational methods

have been developed to predict strong and weak ties [Gilbert and Karahalios, 2009]. People’s friendship network structure has been found to be informative for predicting romantic relationships [Backstrom and Kleinberg, 2014]. This dissertation continues to leverage the opportunities to study network structure by presenting a large-scale analysis on how social network structure predicts trust in Facebook groups.

Finally, the third wave of the digitalization of exchange — **the digitalization of the exchange of resources** — took place at scale through sharing economy platforms such as Airbnb (founded in 2008); Uber (founded in 2009); and Lyft (founded in 2012). Building on top of the payment and social networks infrastructure developed by the first two waves of the digitalization, the last wave of the digitalization of social exchange requires that people not only trust others with their financial resources, personal information, but also their *physical safety*. When people stay at strangers’ homes (Airbnb), or hop into strangers’ cars (Uber and Lyft), a great amount of risks are being mitigated through interpersonal trust as well as trust in the sharing economy platforms. One of the most distinct characteristics of these platforms is that they blur the line of social and economical exchange [Hamari et al., 2016, Ikkala and Lampinen, 2015, Lampinen and Cheshire, 2016], For example, in the context of Airbnb, “guests” stay at “hosts” homes paying a fee, constituting an economic transaction. However, unlike a purely hotel-like interaction, some Airbnb hosts also value social interactions as well as potential friendships built through hosting [Lampinen and Cheshire, 2016].

The in-between nature of sharing economy interactions creates new questions about how people trust each other on these platforms. Social mechanisms such

as homophily can play a role in establishing trust [Abrahamo et al., 2017]. The problem of bias can also arise [Edelman and Luca, 2014]. At the same time, reputation systems can help offset homophily tendencies [Abrahamo et al., 2017]. Other mechanisms such as self-disclosure can also establish trust by reducing uncertainty [Berger and Calabrese, 1975, Ellison and Hancock, 2013]. This dissertation presents a case study on self-disclosure and trust through computational analysis of the language in host profiles on Airbnb. As sharing economy continues to evolve and mature, new questions about trust will continue to arise. For example, in addition to lodging, Airbnb has also launched “experiences”¹, where hosts lead in-person tours for guests. “Experiences” require higher levels of interpersonal interaction and hence also higher levels of trust. Understanding how trust functions at different levels and how to predict trust computationally by level remain future work.

Taken together, three waves of the digitalization of exchange have fundamentally **altered the social structure under which people conduct exchanges** — creating new questions about how people trust each other within these new structures. My work address some of these new questions about trust in the context of each wave, with a focus on leveraging computational methods to build algorithms to understand and predict trust. Through computational methods on analyzing images, language, and social networks, we gain a deeper understanding of trust in digitalized social exchange that are perhaps hard to obtain with qualitative or smaller scale studies. The insights on how people trust each other in each context can inform the design of the digital platforms facilitating exchange. At the same time, these new computational models also have the potential to be built directly into the digital platforms to influence the ranking

¹<https://www.airbnb.com/help/article/1581/what-are-experiences>

and recommendation of potential exchange partners, fostering trust through engineering.

More importantly, the work presented in this dissertation prompts us to think about **trust in digitalized exchange** differently. Originally, trust as a concept was developed as a social construct *between people* [Luhmann, 1979]. The concept of *online trust*, or *eTrust*, gained importance with the first wave of digitalization of exchange in the context of trust on online commerce platforms [Cook et al., 2009]. However, as discussed in the contexts of SNSs, trust is also affected by networks. In addition, increasingly, algorithms are gaining more control over how we trust online through personalization and recommendations. Online trust as a term or framework for thinking about trust in these socio-technical systems is becoming insufficient.

To incorporate and emphasize the roles of networks and algorithms in trust in digital platforms that facilitate exchange, I propose the framework of “**networked trust**”. The term “networked” is borrowed from the concept of “networked individualism”, which refers to the new social structure where people are embedded in social networks with more weak ties as a result of digitalization of social relationships [Rainie and Wellman, 2012]. There are three focuses in *networked trust*: (1) *cues* in Computer-Mediated Communication; (2) embeddedness in social *networks*; and (3) increasing mediation by *algorithms*.

The first focus of networked trust, **cues in Computer-Mediated Communication (CMC)**, refers to how cues that are available in digitalized social exchange differ from cues in face-to-face interactions. This focus of networked trust is most aligned with the traditional meaning of the term “online trust”. Increased uncertainty brought by CMC results in the need for overcoming such uncertainty

to establish trust. In my work, I discuss how images and language cues play a role in establishing trust in Chapter 3 and Chapter 4, by leveraging opportunities brought by new computational methods to deepen our understanding of cues.

The second focus of networked trust, **embeddedness in social networks**, refers to how information about social networks can affect trust. For example, on LinkedIn, when users view another person's profile, LinkedIn informs the users about how they are connected in the LinkedIn network. Information such as the number of common connections, the degree of separation, the specific people that can act as a bridge to introduce the users, are important for establishing trust in the context of networking. In my work, Chapter 5 shows how networks contribute trust in the context of social groups on Facebook.

Finally, the third focus of networked trust, **increasing mediation by algorithms**, refers to how trust can be impacted by algorithmic reconfiguration of social exchange relationships. It is no secret that modern digitalized exchange platforms use algorithms for personalization [Gubin et al., 2017, Bakshy et al., 2015, Grbovic and Cheng, 2018, Linden et al., 2003, Zhou et al., 2008]. Notable examples of such algorithmic mediation include, Facebook News Feed ranking [Gubin et al., 2017, Bakshy et al., 2015], Netflix Prize [Zhou et al., 2008], and Amazon's item-to-item collaborative filtering [Linden et al., 2003]. Increasingly, sharing economy platforms such as Airbnb also deploy algorithmic ranking and matching algorithms to recommend exchange partners, potentially optimizing for conversion [Grbovic and Cheng, 2018]. A key part of the utility value of Uber and Lyft comes from the algorithmic matching of drivers and passengers through dispatch and routing. These algorithms create efficiency in exchange relationships, but can also create problems for trust. The opaqueness and invis-

bility of such algorithms can hinder not only how users trust the system [Eslami et al., 2015], but also how people *trust each other* when mediated by algorithms. The algorithmic mediation in interpersonal exchange also raises new questions about bias, which can further reduce trust in the platform. For example, should matching algorithms on online dating platforms recommend potential partners for users of their own race [Hutson et al., 2018]? This dissertation does not directly address this focus of networked trust, but Chapter 6 discusses how future work can make an important theoretical contribution through the development of AI-Mediated Exchange Theory.

The work presented in this dissertation sets the foundation for several areas of **future work**. The networked trust lens can be directly applied to new contexts to understand how different factors affect trust in a networked environment. For example, in the context of **misinformation**, networked trust framework will look at how factors work together to impact trust in information, including cues (the credibility of news sources), networks (how information spreads through social networks), and algorithms (the impact of algorithmic curation).

At the same time, networked trust can be extended to chart two future directions of research: **AI-Mediated Communication (AI-MC)** and **AI-Mediated Exchange Theory (AI-MET)**. The first focus of networked trust, cues in Computer-Mediated Communication, can be extended to **AI-Mediated Communication (AI-MC)**. In interpersonal communication, increasingly AI-powered systems can modify, augment, and even generate cues in self-presentation or messages to optimize for interpersonal outcomes. For example, the algorithms built in this dissertation predicting trust in Airbnb host profiles can be fitted to improve a given profile to make it appear to be more trustworthy. AI-MC aims to under-

stand how such AI-powered systems might impact interpersonal communication outcomes, which is important for the future of networked trust. Early studies have already observed initial effects of AI-MC [Jakesch et al., 2019] while exact definition and research agenda are being developed and completed [Naaman et al., 2019].

At the same time, the third focus of networked trust, increasing mediation by algorithms, can be extended to **AI-Mediated Exchange Theory (AI-MET)**. AI-MET is an extension of social exchange theory [Cheshire, 2007, Yamagishi and Cook, 1993, Bearman, 1997]. By viewing AI systems as mediating social exchange in socio-technical systems, AI-MET articulates different mechanisms through which AI reconfigures social exchange relationships. AI-MET can allow different perspectives in exploring the relationships between humans and AI to speak to each other. AI-MET can be useful in understanding issues of AI in socio-technical systems such as how trust in the algorithms affects trust in another person when their relationships are mediated by algorithms.

1.1 Outline

The rest of this dissertation is organized as such: Chapter 2 first sets a theoretical foundation for the definition of trust using social exchange theory. Chapter 3 - Chapter 5 each details a case study on the computational understanding and prediction of trust. Each case study investigate trust in the context of a platform representing specific wave of the digitalization of social exchange, including peer-to-peer marketplaces, social networks, and sharing economy platforms. Then, Chapter 6 charts future research agenda based on the work presented in

this dissertation.

Specifically, Chapter 2 reviews the literature that my work builds upon. The first section reviews social exchange theory as the theoretical foundation for the definition of trust. Through social exchange theory, we view the world as an equilibrium established through negotiated exchange. Trust mitigate risks during exchange, leading to higher efficiency and better societal outcomes. Then, key literature on the determinants, outcomes, and the measurement of trust was reviewed through a multilevel perspective. Special attention was paid to literature in online trust, and the lens of networked trust was discussed in more detail in relation to literature in online trust. Finally, literature related to the flip side of trust, bias was reviewed, including the racial, gender, status, and algorithmic bias.

In the two chapters that follow, I cover work on the first focus of networked trust — how trust is established through images (Chapter 3) and language (Chapter 4) cues.

Chapter 3 examines the role of image cues on trust in the context of online peer-to-peer buy-and-sell marketplaces. Using techniques from computer vision, including both image feature-based and deep learning, I show that we can predict image quality with high accuracy (close to 90%), and that high quality images predicted using our model outperform stock imagery in generating trust in marketplaces. Finally, I also show that higher quality images predicted by our model leads to higher sales, showing how algorithms predicting trust translate to real-world outcomes.

Chapter 4 discusses how we can learn about trust on Airbnb based on lan-

guage cues in host profiles. There are two parts in this chapter. First, I discuss how we constructed the dataset of Airbnb host profiles with perceived trustworthiness, while developing different topics that are frequently mentioned in the profiles. Then, I present a computational framework to predict trust in host profiles leveraging natural language processing and machine learning.

Next, Chapter 5 expands on the second focus of networked trust — trust in social exchange that is embedded in social networks. Chapter 5 presents a large-scale comprehensive study of trust in people’s social groups on Facebook. Leveraging survey data with social network and interactions data, we learn the patterns of trust across a diverse range of groups of different sizes and categories. I show how both individual characteristics and group characteristics such as group network size and density matter for trust in groups. In addition, I show how trust in groups may lead to different group outcomes, and how we can leverage these findings to design for better online groups.

Finally, Chapter 6 charts three directions for future work: (1) networked trust and misinformation; (2) AI-Mediated Communication (AI-MC); and (3) AI-Mediated Exchange Theory (AI-MET).

1.2 Summary of Contribution

This dissertation contributes to the field of Human Computer Interaction (HCI), Computer-Supported Cooperative Work and Social Computing (CSCW), and Computational Social Science in the following ways:

1. Through three comprehensive studies of computational trust in different

contexts, this dissertation extends previous understanding of trust leveraging new techniques (e.g., deep learning), new contexts (e.g., sharing economy), and new data sources (e.g., network structure). Each study showed that not only we were able to understand and predict trust using a variety of different algorithms, but also trust has real-world impact on specific outcomes. Importantly, all work presented in this dissertation is based on large-scale real-world systems — covering major categories of social exchange platforms that represent different waves of digitalization of social exchange (online commerce, social networks, and sharing economy). The findings in this dissertation validate and extend what we know about trust, and can inform the design of future digital platforms to better facilitate trust.

2. This dissertation proposes a framework of “networked trust”, which can be a useful lens to further the discussion from “online trust” to focus more on factors that are increasingly important — networks and algorithms. Networked trust can be applied to different contexts where trust is urgently needed, such as misinformation, to understand how cues, networks, and algorithms work together to affect trust.
3. Finally, networked trust can further be extended to chart two important research agenda for future development — AI-Mediated Communication and AI-Mediated Exchange Theory. AI-Mediated Communication focuses on extending the cues focus of networked trust, by better understanding the impact of AI on interpersonal outcomes when cues are subject to the modification, augmentation, and generation by AI. AI-Mediated Exchange Theory, on the other hand, focuses on incorporating the role of algorithms in networked social exchange broadly. Early work has been established

in AI-Mediated Communication [Jakesch et al., 2019, Naaman et al., 2019], while AI-Mediated Exchange Theory remains future work.

CHAPTER 2

BACKGROUND

Trust is the chicken soup of social life. [...] Like chicken soup, trust appears to work somewhat mysteriously.

Eric M. Uslaner

*in *Producing and Consuming Trust* (2000)*

The work presented in this dissertation is highly inter-disciplinary. My definition and the scoping of trust build on previous scholarship across disciplines, just to name a few, sociology [Luhmann, 1979, Gambetta, 1988], psychology [Rotter, 1967], economics [Berg et al., 1995], business [Grabner-Kraeuter, 2002], and political science [Miller, 1974]. This chapter reviews prior scholarship on trust across disciplines that my work builds on through a multilevel perspective.

One key challenge of researching trust is that the term is heavily loaded with many different meanings and bears a great deal of ambiguity. To clarify the concept being studied, this chapter first provides a theoretical framework under which we define trust, heavily drawing from social exchange theory (Section 2.1). Then, in order to talk about trust concretely in specific contexts, I review various measurements, determinants and outcomes of trust in prior literature (Section 2.2). The work presented in this dissertation draws from the definition and measurements reviewed here. In addition, because this dissertation focuses on trust in digitalized exchange, I devote a section to specifically focus on previous literature on online trust (Section 2.3). As networks and algorithms play a more important role in digitalized exchange, I propose the framework of

“networked trust” as a lens to reason about trust in networked environments and to address new challenges that arise.

Finally, trust as a mechanism to facilitate exchange has key limitations. Trusting someone usually means distrusting others. As the process of forming trust can be subjective, bias can occur when there is need to trust. Section 2.4 reviews literature on four different types of biases, racial, gender, status, and algorithmic. Bias can be manifested in algorithms in subtle ways. Before deploying computational models that predict trust to real-world applications, we need to be cautious and aware of the potential biases that these algorithms may have.

2.1 Theoretical Foundations of Trust

Trust is an inherently social phenomenon. To research trust and to connect the results meaningfully to societal outcomes, it is important to view it not in isolation but rather grounded in broad sociological traditions and structures. This section provides the theoretical foundation for defining trust — drawing from social exchange theory in sociology.

Sociology, the study of society, has multiple paradigms, which are “fundamental images of the subject matter within a science” [Ritzer, 1975]. Paradigms are useful to “define what should be studied, what questions should be asked, how they should be asked, and what rules should be followed in interpreting the answer obtained” [Ritzer, 1975]. According to Ritzer, major paradigms in sociology include: structural functionalism (or systems theory), conflict theory, symbolic interactionism, and social exchange theory [Ritzer, 1975]. Structural functionalism is oriented to analyze social structures and institutions, viewing

individuals as largely controlled by the structures they are embedded in. Conflict theory, on the other hand, views the world in constant conflicts and focuses on studying disintegration and change in society. Symbolic interactionism places greater emphasis on the mental process of the individuals as active creators of social reality. Finally, social exchange theory views social change and stability as a process of negotiated exchanges between different parties, primarily driven by the process of reinforcement [Emerson, 1976].¹

We can view trust under any of these paradigms, and the resulted research would have different focuses. For example, trust when viewed under structural functionalism serves as a “mechanism to reduce social complexity” [Luhmann, 1979].² Such a viewpoint can be found in one of the earliest classic work on trust: the 1979 book *Trust and Power* by Niklas Luhmann [Luhmann, 1979]. The complexity of social systems is due to both the scale as well as uncertainties in interacting with each agent in the system. Analyzing the motivation and potential behaviors of each agent is cognitively impossible. Trust provides a shortcut for decision-making when navigating social systems and thus reduces social complexity. As the “lubricant of social interaction” [Arrow, 1974], trust facilitates social interactions, reducing the whole system’s complexity. On the other hand, if we view trust under conflict theory, we can focus more on the dynamics of distrust and mistrust [Kramer, 1999, Ely, 1980, Marsh and Dibben, 2005, Kioussis, 2001]. Trust under symbolic interactionism would include work that focuses on the psychological process such as how nonverbal cues affect trust in partners [Lee et al., 2013, DeSteno et al., 2012]. Finally, trust under

¹We do not expand on the limitations of each paradigm in this dissertation as it is not the focus, but curious readers can find relevant discussions in [Ritzer, 1975, Emerson, 1976].

²It is worth noting that trust is not the only mechanism to reduce social complexity. Power and control can also reduce social complexity by reducing the freedom of others, through possibly coercion, contracts, and reputation systems [Luhmann, 1979].

social exchange theory would focus on how trust enables cooperation — as seen in another classic book on trust, *Trust: Making and Breaking Cooperative Relations* [Gambetta, 1988].

In this dissertation, I draw primarily on the last paradigm, social exchange theory. Viewing trust under social exchange theory has the following advantages: (1) It allows for a clean definition of trust in the context of risks in exchange; (2) It allows for a multilevel view of trust as social exchange theory accounts for both micro and macro exchanges; (3) It allows for the easy incorporation of digital platforms in exchange. We expand on each point below.

First, as mentioned above, trust is a loaded term, used colloquially to refer to a variety of different things. Grounding trust in social exchange theory allows for a clean definition of trust in the exchange context that mitigates risks. Social exchange theory views social change and stability as a process of negotiated exchanges between different parties. Social relations are the units of analysis [Emerson, 1976] in social exchange theory. By analyzing costs and rewards in these relations, social exchange theory reasons about how people make decisions in the social world quasi-economically. In a potential exchange context, risks can create obstacles of collaboration, preventing the exchange from taking place. Trust, *a decision to be vulnerable to the other party involved in the exchange*, can mitigate risks and thus enable exchange and cooperation.

As Diego Gambetta defined:

“Trust (or, symmetrically, distrust) is a particular level of the subjective probability with which an agent assesses that another agent or group of agents will perform a particular action, both before he

can monitor such action (or independently of his capacity ever to be able to monitor it) and in a context in which it affects his own action. When we say that the probability that he will perform an action that is beneficial or at least not detrimental to us is high enough for us to consider engaging in some form of cooperation with him. Correspondingly, when we say that someone is untrustworthy, we imply that the probability is low enough for us to refrain from doing so.” [Gambetta, 1988]

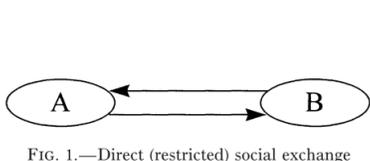


FIG. 1.—Direct (restricted) social exchange

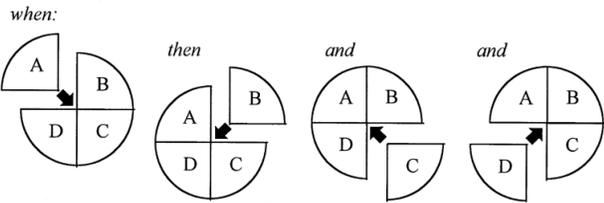


FIG. 2.—Indirect (generalized) social exchange: net form

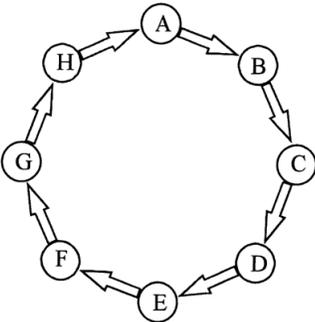


FIG. 3.—Indirect (generalized) social exchange: chain form

Figure 2.1: Direct and two forms of generalized (indirect) exchange illustrated by Bearman 1997. Each character represents a party involved in the exchange. In direct social exchange, there is direct transfer of value or goods between A and B. In generalized exchange, multiple parties pool their resources together to produce greater value, which is then further distributed.

Second, unlike pure macro-focused theory such as structural functionalism, social exchange theory focuses both on direct (micro) and generalized (macro) exchange [Emerson, 1976]. Direct exchange usually refers to dyadic exchange,

where two parties engage in a mutually beneficial exchange relationship, transferring value or goods directly between two parties. At the same time, social exchange theory also discusses generalized exchange — or sometimes referred to as “productive exchange” — the process where resources from multiple parties are pooled together to produce greater value, which is then further be distributed [Emerson, 1976, Bearman, 1997, Yamagishi and Cook, 1993]. Figure 2.1 contains illustrations of direct exchange (top left), and two different forms of generalized exchange (top right: net form, bottom: chain form) [Bearman, 1997]. We can view trust in the context of facilitating both direct and generalized exchange — leading to a multilevel view of trust that we expand in Section 2.2. The multilevel view of trust refers to the study of trust at different levels of abstractions, for example, trust at the individual level, trust in groups, trust in organizations, trust in the government, and most broadly, trust in the society. A multilevel view of trust is useful for organizing prior work on trust across multiple disciplines and traditions. At the same time, it can also inform the development of computational models that incorporate predictors on both micro and macro levels.

Finally, the exchange view of trust allows for easy incorporation of the effects of online platforms on social exchange and trust. The digital platforms alter the social structure under which people conduct exchange, creating new questions about trust during the process. In the past few decades, our social exchange structure has undergone fundamental changes through three waves of the digitalization of exchange: (1) the digitization of the exchange of goods; (2) the digitization of social relationships; and (3) the digitalization of the exchange of resources through sharing economy platforms. As a result, people rely on cues that are available online such as images and language to establish trust, rather

than relying on nonverbal cues in face-to-face interactions. In addition, how people are embedded in social networks plays a key role in establishing trust. Defining trust under the framework of social exchange theory also allows for the easy incorporation of networks as factors that impact trust in mediated exchange. In fact, the “networked trust” framework proposed by this dissertation is such an extension from social exchange theory [Cook and Whitmeyer, 1992].

In conclusion, the work presented in this dissertation views trust through the lens of social exchange theory, one of the many sociological paradigms. Throughout this dissertation, trust at the highest level of abstraction is defined as: **a willingness to make oneself vulnerable to other parties in direct or generalized exchange, which mitigates risks and enables exchange and cooperation.** However, in specific contexts, trust can be defined and measured in more context-specific ways. We review literature on measuring trust next.

2.2 Measuring Trust

When I tell people that I conduct research on trust, the first question they usually ask is, “Well, but how do you measure trust?” It is a fair question. The first step to research anything is to define and measure it, especially for a loaded term such as trust. This section focuses on the measurement of trust through first providing a categorization of common definitions of trust, and then reviewing literature on three key methods: (1) surveys; (2) real-world behavior proxies; and (3) lab-based trust games. In particular, I use a multilevel perspective to categorize common definitions of trust to provide some clarity on what trust means across different disciplines and traditions. In addition, as I show in Chapter 5, a multilevel

perspective of trust can inform the development of computational models by incorporating both micro and macro level factors to improve prediction accuracy.

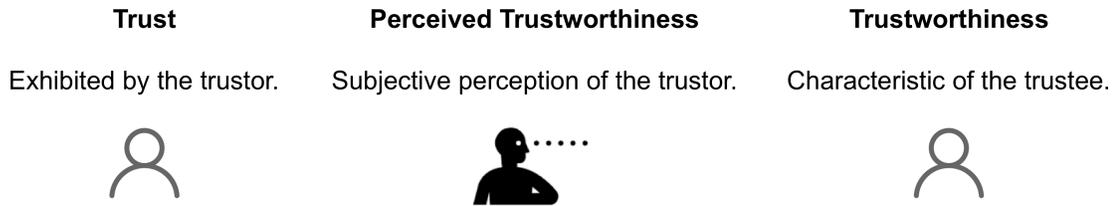


Figure 2.2: Differences among trust, trustworthiness and perceived trustworthiness.

Before we delve in, it is helpful to first clarify a few closely related concepts: *trust*, *trustworthiness*, and *perceived trustworthiness*. Hardin [Hardin, 2002] and Kiyonari et al. [Kiyonari et al., 2006] pointed out that the general literature on trust typically overlooked the difference between trust and trustworthiness, and used the term interchangeably. As illustrated in Figure 2.2, *trust* is exhibited by a trustor — a decision to be vulnerable, and *trustworthiness* is a characteristic of the trustee [Kiyonari et al., 2006]. In particular, there are also important differences between trustworthiness and *perceived* trustworthiness. Trustworthiness refers to whether the trustee behaves in a way that is not harmful to the trustor in reality even though that it is well within the trustee’s capacity and freedom to do so [Kiyonari et al., 2006]. Perceived trustworthiness, on the other hand, is in the eyes of the trustor. It refers to a subjective belief of the trustor about the trustworthiness of the trustee, which can be of course be subjective and biased [Cheshire, 2011].

The relationships among three concepts are complex. In one-shot interactions, trust does not necessarily beget trustworthiness [Kiyonari et al., 2006]. Perceived trustworthiness does not necessarily lead to trust either. The decision to trust might be mediated by the risk tolerance of the trustor. In this dissertation, when

necessary, I try to distinguish between trust, trustworthiness, and perceived trustworthiness. Such distinction is helpful for measuring of trust. But when I discuss general concepts, such as in the context of “networked trust”, for simplicity, I use “trust” to encapsulate trust, trustworthiness, and perceived trustworthiness.

2.2.1 Determinants and Outcomes of Trust — A Multilevel Perspective

As a concept central to social exchange, trust has received research attention from a variety of disciplines in the past, just to name a few, sociology [Luhmann, 1979, Gambetta, 1988], psychology [Rotter, 1967], economics [Berg et al., 1995], business [Grabner-Kraeuter, 2002], and political science [Miller, 1974]. Different traditions, focuses, and levels of abstractions across disciplines make it difficult to reach an agreement on how trust should be measured, what leads to trust, and its outcomes. To bridge the gaps among different disciplines, I provide a multilevel perspective of prior literature on trust, focusing on the determinants and outcomes of trust first, and its measurement in Section 2.2.2.

What is a multilevel view of trust? When people talk about trust, they can be referring a variety of different things. Most commonly, trust refers to an individual’s trust in other entities across levels of abstractions. As depicted in Figure 2.3, trust can be discussed in the context of trust in another person, trust in several other people within a group or organizations, trust in a specific company or platform, or trust in the society in general.³ Below I review specific work on

³One can of course theoretically discuss trust originating from not the individual, but other

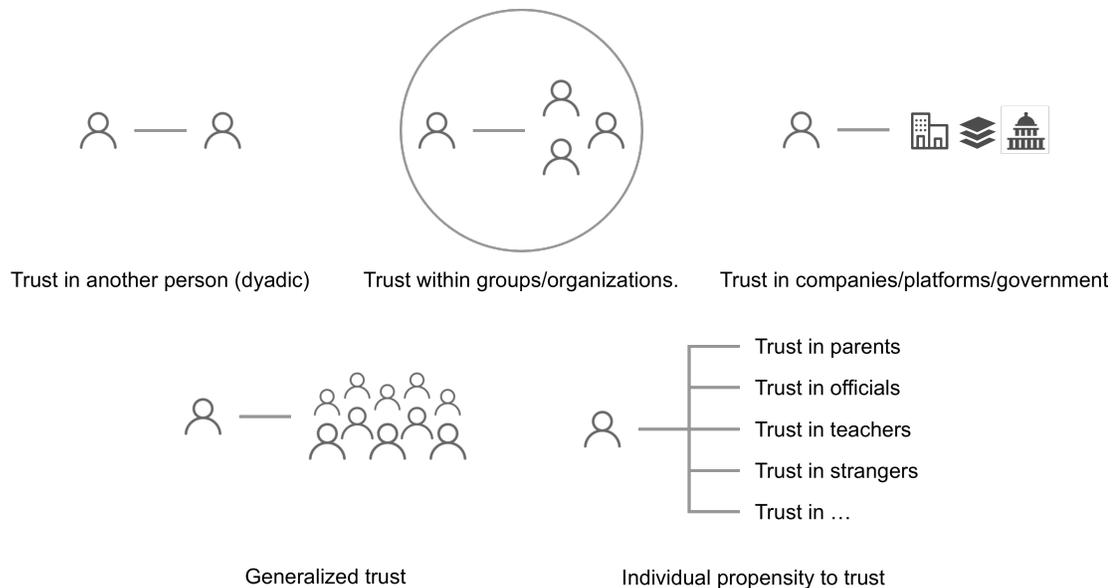


Figure 2.3: Multilevel perspective of trust.

trust at each level of abstraction.

Trust in another person (dyadic). Trust in another person, or commonly referred to as dyadic trust, is one of the most common conceptualizations of interpersonal trust [Larzelere and Huston, 1980, Berg et al., 1995]. Dyadic trust refers to the decision to trust or the assessment of trustworthiness of a specific individual, as opposed to people in aggregate [Larzelere and Huston, 1980]. From a social exchange theory point of view, dyadic trust is trust involved in direct exchange. One key advantage of studying dyadic trust is that it is the “minimal” meaningful unit of interaction that involves trust, providing conceptual simplicity to investigate factors that affect trust.

The field of social psychology and economics both discuss dyadic trust heavily, but using different methods. In psychology, dyadic trust is most commonly measured with techniques in psychometrics, by developing interper-

entities, such as how a company trusts another company. However, such discussion is less common in prior literature and also not the focus of this dissertation.

sonal trust scales. For example, the *Dyadic Trust Scale* measures dyadic trust with agree/disagree items including “My partner is primarily interested in his (her) own welfare”, and “My partner is perfectly honest and truthful with me” [Larzelere and Huston, 1980]. Such measures of dyadic trust has shown that dyadic trust correlates with love and intimacy of self-disclosure [Larzelere and Huston, 1980]. The work presented in Chapter 4 in this dissertation follows this tradition and investigates dyadic trust on Airbnb using a interpersonal trust scale adapted to the domain of lodging.

In economics, dyadic trust is most commonly measured through trust games [Berg et al., 1995]. Initially developed by Berg et al., the trust game is an economic decision-making game that has built-in risk dynamics to allow for the measurement of trust and trustworthiness [Berg et al., 1995] (I expand on details of the trust game in Section 2.2.2). Since then, the trust game has become a useful instrument to study trust in dyadic settings in and outside of the field of economics, and often in experiments [Kiyonari et al., 2006, Kuwabara, 2015, Abrahao et al., 2017, Yamagishi et al., 2009, Resnick and Zeckhauser, 2002, Qiu et al., 2018, Merrill and Cheshire, 2017]. For example, Kuwabara found that reputation systems can reinforce trust and trustworthiness through experiments using trust games [Kuwabara, 2015]. In another experiment, it was found that reputation can overcome homophily in establishing trust, again using trust games to measure dyadic trust [Abrahao et al., 2017]. Finally, in an investigation on how bio signals affect trust, trust games were used to show that elevated heartrates reduce trust [Merrill and Cheshire, 2017].

Trust within groups/organizations. While it is useful to study trust in direct exchange (dyadic trust), trust in generalized exchange is also important to under-

stand. One common form of generalized social exchange is through social groups or organizations. Trust within groups or organizations⁴ refers to trust in people in aggregate in the group or organization, as opposed to a specific individual. In particular, my work in Chapter 5 focuses on predicting trust within groups quantitatively.

Work in social psychology, communication, and management science has examined trust within groups or organizations. In social psychology, trust within groups was found to stem from basic properties of the group such as its size [Brewer, 1991]. For instance, experiments have shown that people identify more strongly with smaller groups [Simon and Brown, 1987]. Trust within groups was also found to indirectly influence group task performance [Dirks and Ferrin, 2002]. In communication, trust within groups was found to correlate with rule following behaviors [Walther and Bunz, 2005]. However, trust within groups can also be fragile and temporal when the group is formed around a common task with a finite life span [Jarvenpaa and Leidner, 1999, Meyerson et al., 1996].

In management science, an important theoretical framework on organizational trust was proposed by Mayer et al. in 1995, which took an integrative approach by synthesizing prior work on trust from multiple disciplines [Mayer et al., 1995, Schoorman et al., 2007]. Three dimensions of trust, ability, benevolence, and integrity, were proposed. Ability refers to the competency or technical skills of the trustee; benevolence refers to the extent the trustee is believed to want to do good to the trustor; and integrity refers to the belief that the trustee will adhere to a set of principles that the trustor finds acceptable [Mayer et al.,

⁴I use the term trust *within* groups or organizations here to distinguish better from trust in companies/platforms/government below, where the trustor is not part of the other entity. However, in Chapter 5, the term trust *in* groups is used to refer to trust within groups for simplicity.

1995]. Though originally developed in the organizational context, in my work I adapt these dimensions in a variety of settings to measure trust in surveys, including trust in platforms in Chapter 3, dyadic trust in Chapter 4, and trust within groups in Chapter 5. Using such framework, trust in management within organizations has been shown to positively relate to employee's ability to focus attention on value-producing activities [Mayer and Gavin, 2005]. In addition, trust within teams has been found to be positively related to perceived task performance, team satisfaction, relationship commitment, and negatively related to stress and tardiness [Costa et al., 2001, Bijlsma and Koopman, 2003, Colquitt et al., 2007].

Finally, as people more commonly engage with social groups online, new opportunities arise to study trust within groups online, especially how group network structure relates to trust. Several studies in Human-Computer Interaction began to examine trust in social groups in social networks. In one qualitative research on Facebook buy-and-sell groups, trust was found to be fostered through mechanisms such as closed membership and sanctioning [Moser et al., 2017]. In addition, network density was also found to lead to higher trust in buy-and-sell groups [Holtz et al., 2017]. My work builds on existing work to understand and predict trust within Facebook groups in Chapter 5.

Trust in companies/platforms/government. Another important level of trust is trust in companies, platforms, and the government. In contrast to trust within groups or organizations where the trustor is part of the group, the trustor here is not a part of the organization.

Most commonly, marketing literature discusses trust in companies and brands, while political science literature discusses trust in the government. In

marketing, trust in brand was found to determine consumer satisfaction and loyalty, both in terms of purchase behavior and attitudes [Chaudhuri and Holbrook, 2001, Delgado-Ballester and Luis Munuera-Alemán, 2001]. In addition, consumer trust in online commerce platforms has also received research attention, especially in the early 2000s when they first emerged [Corritore et al., 2001, Gefen and Straub, 2004]. In my work, Chapter 3 presents a study of trust in online commerce platforms. In political science, trust in government, or public trust, is a crucial research question especially important for the study of democracy [Hardin, 1999, Chanley et al., 2000]. Prominent political scientists such as Francis Fukuyama argue that public trust fosters social and economic prosperity [Fukuyama, 1995]. However, in past few decades, literature has tracked a steady decline of generalized trust in society [Pew, 2019, Miller, 1974, Chanley et al., 2000]. Relatedly, trust in media and news has also been on decline, especially as concerns over misinformation grow [Kohring and Matthes, 2007, Gunther, 1988, Allcott et al., 2019, Flintham et al., 2018, Grinberg et al., 2019, Flintham et al., 2018].

Finally, as technological platforms play an increasingly important role in every aspect of people's lives, trust, or rather, distrust in technological platforms has gained significant attention, especially in the areas of Human-Computer Interaction, Computer Supportive Collaborative Work, Science and Technology Studies, and law and policy research. In particular, distrust in social networking sites such as Facebook has brewed over the years, exacerbated by the emotional contagion study [Kramer et al., 2014], and Cambridge Analytica scandal⁵ [Dwyer et al., 2007, Fogel and Nehmad, 2009, Lankton and McKnight, 2011]. Opacity and bias in algorithms that are ubiquitous on digital platforms are likely to

⁵<https://www.nytimes.com/2018/03/17/us/politics/cambridge-analytica-trump-campaign.html>

further erode trust in platforms [Barocas and Selbst, 2016, Buolamwini and Gebru, 2018, Datta et al., 2018, Bolukbasi et al., 2016, Eckhouse et al., 2019].

Taken together, the decline of public trust, trust in media and news organizations, and distrust in technological platforms, are among the major challenges that the world faces around trust today. It is important for future research on trust to address these challenges.

Generalized trust and individual propensity to trust. Trust in the broadest sense measures trust in other people in the society in general and in aggregate. Trust in this context is commonly referred to as “generalized trust” [Nannestad, 2008], a “propensity to trust” [Ferguson and Peterson, 2015], or a “disposition to trust” [Wu et al., 2010].⁶ Generalized trust is considered as a stable individual trait, similar to that of personality [Rotter, 1967]. Previous work has also studied cross-country differences in generalized trust, where social polarization in the form of income inequality and ethnic diversity was found to reduce trust [Bjørnskov, 2007]. Therefore, it is useful as a control variable in the study of trust in other contexts. For example, in Chapter 5, generalized trust was used as a control variable in the prediction of trust within groups and was found to have significant predictive value.

Exact definitions of generalized trust and individual propensity to trust are slightly different, although they are largely considered to be very similar. Generalized trust is usually measured with the trust question in General Social Survey (GSS) or the World Values Survey (WVS), “Generally speaking, would you say that most people can be trusted or that you can’t be too careful in dealing with people?” Such measure has been subjected to criticism that that the interpretation

⁶Technically, trust in this context refers to the average level of perceived trustworthiness of others in the society.

of “most people” could vary widely person to person, therefore can be unreliable with stochastic errors [Nannestad, 2008]. In addition, generalized trust has been shown to have little predictive power when it comes to specific individual behaviors in games that involve trust [Glaeser et al., 2000]. On the other hand, individual propensity to trust has followed a psychometric tradition, measured with instruments such as the Rotter Interpersonal Trust Scale (ITS) [Rotter, 1967]. The 25 agree/disagree items in ITS ask about perceived trustworthiness across a variety of different roles in society, and take the average of the responses as the trust composite score. These agree/disagree items include “Parents can usually be relied upon to keep their promises.”, “Most elected public officials are really sincere in their campaign promises”, etc.

One line of work suggests that individual propensity to trust is rooted in life experiences and societal norms [Rotter, 1971, Bowlby, 1969, Ainsworth et al., 2015]. Propensity to trust is considered as an important variable in the development of healthy family relationships and personalities in children [Rotter, 1967] It has also been associated with being older, married, having higher family income and college education and living in a rural area, but not with gender [Taylor et al., 2007, Paxton, 2007]. A propensity to trust is also related to other personal traits, such as risk-taking [Cook et al., 2005], feelings of social support, and group loyalty [Barrera Jr and Ainlay, 1983, Van Vugt and Hart, 2004].

A multilevel perspective. As reviewed above, when we talk about trust, it can refer to trust in entities across different levels of abstractions. Different levels of abstractions, as well as different traditions and language of disciplines created challenges for trust research. To get a holistic view of trust, it is important to synthesize various viewpoints and examine how trust at different levels of

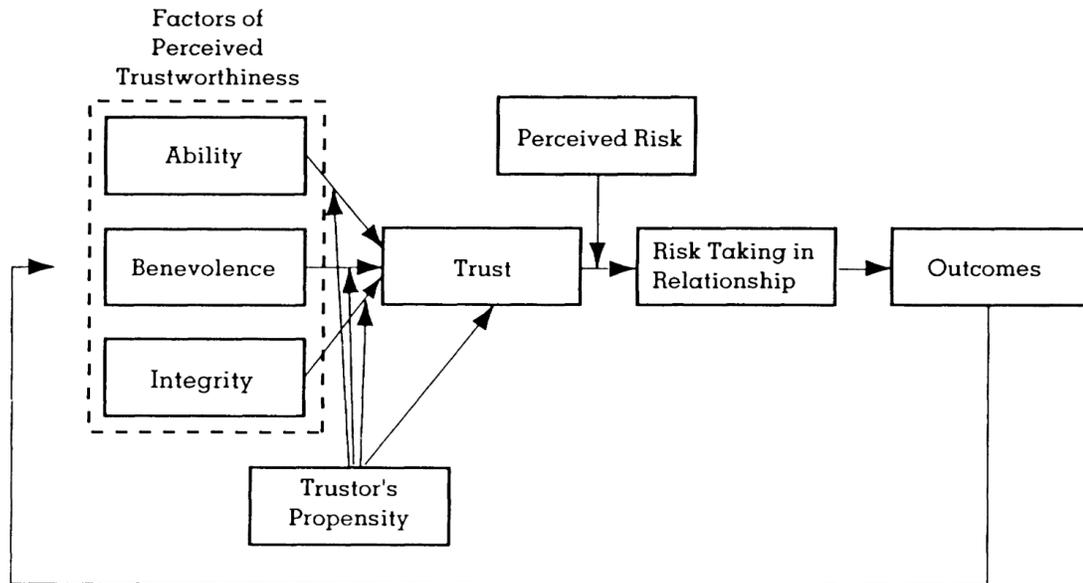


Figure 2.4: The integrative organizational trust model proposed by Mayer et al. 1995

abstractions interact with each other. For example, how does trust in platforms affects trust in information shared by another person?

Mayer et al. 1995 proposed an integrative approach to study trust, which has been highly impactful [Mayer et al., 1995]. In particular, as shown in Figure 2.4, Mayer et al. emphasized how individual propensity to trust can mediate the relationship between impressions of perceived trustworthiness and the decision to trust.

At the same time, with increasing digital mediation in our social exchange, the integrative model proposed by Mayer et al. needs further development. Interpersonal exchange takes place embedded in social networks (e.g., Facebook, LinkedIn, etc.). In addition, increasing algorithmic mediation in online platforms is also not accounted for in Mayer et al.'s model. A new framework is needed to address the challenges in trust brought by social networks and algorithmic

mediation. To do so, in Section 2.3, I develop a framework of “**networked trust**” to incorporate these new factors in digitalized exchange.

2.2.2 Methodological Note on Measuring Trust

When developing new research projects on trust, deciding how to measure trust is often challenging. Here I offer a brief methodological note on the measurement of trust to facilitate future research. There are mainly three ways that prior literature used to measure trust: (1) surveys; (2) real-world behavior proxies; and (3) lab-based trust games. Each method has strengths and weaknesses, which we discuss briefly below.

Surveys are among the most widely used instrument to measure trust. We can use surveys to ask the willingness for the trustor to make a trust decision, though usually in a hypothetical setup. For example, Chapter 4 used a survey-based experiment to ask about the decision to stay with a particular host on Airbnb. In another example, Kizilcec used a questionnaire to study how transparency of an algorithm affects trust [Kizilcec, 2016]. However, such measures of trust have limited construct validity, as self-report answers in hypothetical decision-making scenarios can be unreliable.

On the other hand, surveys can also measure perceived trustworthiness, which is a subjective assessment and an internal state of the trustor. The World Values Survey has a few survey questions tracking generalized trust (perceived trustworthiness of other people in society) since the 1970s across different countries. One of the most well-known survey question on trust is: “Generally speaking, would you say that most people can be trusted or that you need to be

very careful in dealing with people?" [WVS, 2018]. Respondents can choose one of the options between "Most people can be trusted", "Need to be very careful", and "Don't know". This measure and its variations have also been widely used in political science. For example, surveys on political trust ask "How much of the time do you think you can trust the government in Washington to do what is right?" [Miller, 1974].

Finally, reputation systems can be broadly considered as a survey-based method to establish a proxy measure for trustworthy behaviors. After an exchange has taken place, a survey is issued to exchange partners, asking them to provide feedback and rating towards each other. Then the reputation system aggregates the feedback and ratings and displays them for future interactions. It is important to note that reputation systems have inherent limitations such as bias (racial bias, selection bias, and positivity bias), and are often subject to problems including fake reviews and cold start [Fradkin et al., 2015, Resnick and Zeckhauser, 2002, Zervas et al., 2015, Ott et al., 2011]. A more detailed review of prior work on reputation systems can be found in Section 2.3.

Real-world behavior proxies have higher construct validity in measuring trust compared to survey-based method. For example, when someone decides to fund a crowdsourcing campaign [Mitra and Gilbert, 2014], we can consider the decision or behavior to fund as trust in-situ. Such behavior proxies are becoming increasingly available on digital platforms, such as the decision to meet someone offline through online dating platforms, the decision to book an Airbnb, the decision to purchase something online, etc. These behavior proxy measures are being used in trust research to show the impact of trust on real-world outcomes at scale. Abrahao et al. used actual Airbnb behavior data to validate findings

about reputation and trust [Abrahamo et al., 2017]. In this dissertation, real-world behavior proxies have also been an important part of research findings — the work presented in Chapter 3 leveraged eBay sales data, and the work presented in Chapter 5 leveraged Facebook groups behavioral data [Ma et al., 2019a] such as group densification (a behavior proxy for the decision to trust).

At the same time, behavior proxies can be valuable for addressing problems of bias in online marketplaces and reputation systems. When people make decisions about trust, they rely on information that is available to them to assess the perceived trustworthiness of another party. Bias exists when people make up such impressions (more details on bias are available in Section 2.4). As digital platforms accumulate data about the actual behavior of the trustee (e.g., whether a loan is repaid), behavior proxy data can be used to assess the actual trustworthiness of the trustee, reflecting ground truth rather than impressions. According to signaling theory [Donath, 2007], behavior proxies can be considered as “assessment signals” (something inherent to the signal itself connects it to the quality it indicates) as opposed to “conventional signals” (something not inherently reliable, but are kept so through conventions). There are opportunities to make behavior proxy data available to help address the bias problem in online marketplaces. For example, tracking and displaying the real response time of an Airbnb host, or a business on Facebook messenger can establish trust in the true level of responsiveness, rather than claims.

However, assessing the trustworthiness through behavior proxy also faces challenges. By definition, trustworthiness can only be measured *after* an interaction has taken place, while the decision to trust has to be made *before* the interaction. Sometimes, long periods of time pass before we learn the true behav-

ior of the trustee. For example, in the case of funding startups, it may take at least a few years before the investors learn about the true level of trustworthiness of the team being funded. Second, there is no guarantee that future trustworthiness will be consistent with prior behavior (e.g., Ponzi scheme). Repeated measures can be one way to gain more confidence that the trustee will behave consistently to our assessment. But due to limitations in time, repeated measures can be challenging to obtain. Finally, behavior proxy can be hard to collect or interpret when lacking contextual knowledge. For example, in the case of Airbnb, it is very hard to collect data about the actual condition of the listing, or about whether there are hazards such as hidden cameras inside the listings. In the case of Uber, when the GPS data shows that the driver took a wrong route, is it because the driver is not responsible, due to unpredictable conditions on the road and the driver is reacting responsibly? As such, finding better behavior proxies to measure trust remains future work.

Lab-based trust games are another important instrument to measure trust. Although the work presented in this dissertation relies mostly on surveys and behavior proxies to measure trust, trust games can be useful in studying how certain factors affect trust. Trust games are economic games designed to abstract risks and rewards in real-world exchanges to allow for more controlled study of trust, commonly in experimental setups. By abstracting contextual risks into numerical payoffs, all measures of trust can be reduced to a numerical based decision — making it easy to compare across conditions.

Specifically, the original *trust game* was developed by Berg in 1995 — an investment game that uses the amount of investment to denote the level of trust in the participant's partner [Berg et al., 1995]. The setup of the trust game is

collaboration through risk taking: One participant is given a certain amount of money and can choose an amount between zero and the total amount to give to the partner in the experiment (measures trust). The experimenter will triple the amount that is sent to the other participant, and the other participant can decide how much between zero and the total endowed amount to send back to the first participant (measures trustworthiness). The more one trusts the partner, the more amount in the beginning one should send, as both parties together achieve maximum reward that way. However, the risk of the second participant not returning any amount makes it a prerequisite for the first participant to have trust. Results show that trust indeed exists. In one of the conditions, the first participant sent on average \$5.36 out of \$10, and the other participant returned on average \$6.46 out of \$15 endowed.

Many subsequent research adopted trust games. For example, Kiyonari et al. found that trust does not beget trustworthiness in one-shot games [Kiyonari et al., 2006]. Abrahao et al. used trust games to study the relationship between homophily, reputation and trust [Abrahao et al., 2017]. Finally, Glaeser et al. measured trust and trustworthiness using both the survey-based method, as well as trust games, and found that standard attitudinal survey questions about trust predicts trustworthy behavior in trust games much better than they predict trusting behavior [Glaeser et al., 2000]. However, there are also key limitations to using trust games to measure trust. For example, according to a meta-analysis of 162 trust games across 35 countries [Johnson and Mislin, 2011], behavior in trust games is sensitive to the specific setup of the game. Factors such as the multiplier set by the experimenter, whether the play is with a simulated counterpart, or whether subjects play both roles in the experiment all affect trust.

2.3 Online Trust and Networked Trust

A big focus of this dissertation is to address *new* questions about trust that arise as a result of the digitalization of exchange. In this section, I first review prior work on online trust, primarily focusing *profiles* and *reputation systems*. Then, I briefly trace different waves of the digitalization of exchange and how they affect social structure.

Finally, I propose the framework of “networked trust” to reason about trust in the increasingly connected age, with three focuses: *cues*, *networks*, and *algorithms*.

2.3.1 Online Trust

Online trust is not an entirely new phenomenon — work since early 2000s primarily focused on how two mechanisms produce online trust: (1) profiles; (2) reputation systems.

Profiles are an important part of many online systems [Uski and Lampinen, 2014]. Due to the computer-mediated nature of online systems, profiles provide an identity for users that persists over time and the myriad of interactions on the site [boyd and Ellison, 2007]. In online dating sites, profiles provide self-disclosure that attracts interests of other users, while limiting the risks associated with outright deception [Gibbs et al., 2010, Toma et al., 2008].

One way to understand how profiles establish trustworthiness is the Profile as Promise framework [Ellison and Hancock, 2013]. Profile as Promise uses the notion of a promise to characterize the function of a profile. In this view, the

profile is a psychological contract between the profile holder, and the viewer that future interactions (e.g. with a date, a car driver, or an Airbnb host) will take place with someone who does not differ fundamentally from the person represented in the profile. The notion of a promise has been successfully applied to various contexts that require good faith to operate, including online dating [Ellison and Hancock, 2013] and job hunting [Rousseau and Greller, 1994].

Reputation systems represent another well studied mechanism for online trust. A cornucopia of work on reputation systems, focused on reputation in online commerce platforms, showed the usefulness as well as the limitations of this mechanism in establishing online trust [Jøsang et al., 2007].

At its core, reputation systems provide aggregated information about the trustworthiness of potential exchange partners based on how they have behaved in the past. However, selection bias and positivity bias in reputation systems limit the accuracy and usefulness of reputation systems.

Past research on reputation systems has focused on how reputation systems change people's behavior and their limitations.

In terms of how reputation systems **change people's behavior**, Resnick et al. are among the firsts to study trust and reputation systems [Resnick et al., 2000, Resnick and Zeckhauser, 2002, Resnick et al., 2006]. Through a controlled field experiment, Resnick et al. found sellers with a strong reputation receive a price premium reflecting higher trust on eBay [Resnick et al., 2006]. Kuwabara found that reputation systems can reinforce generalized trust and trustworthiness — recalling one's reputation on eBay made participants behave more trustworthily in the relevant roles [Kuwabara, 2015]. Yamagishi et al. conducted online trading

experiments to see how reputation systems can solve the “lemons problem” for online marketplaces by encouraging more trustworthy behavior [Cook et al., 2009]. A most recent experiment using Airbnb users as participants showed that reputation systems can increase trust between dissimilar users, hence mitigating bias created by homophily [Abrahamo et al., 2017].

In terms of the **limitations of reputation systems**, through empirical analysis on eBay, Resnick et al. found that while only half of the buyers provide feedback (selection bias), the overwhelming majority of feedback is positive (positivity bias) [Resnick and Zeckhauser, 2002]. Similar selection bias and positivity bias were observed on Airbnb [Fradkin et al., 2015, Zervas et al., 2015] and freelancing marketplace oDesk [Filippas et al., 2018]. Non-reviewers on Airbnb were found to have worse experiences than reviewers [Fradkin et al., 2015]. As a result, “every stay is above average” on Airbnb [Zervas et al., 2015] — nearly 95% of Airbnb properties boast an average user-generated rating of either 4.5 or 5 stars (the maximum).

In addition to selection bias and positivity bias, reputation systems are also vulnerable to fake reviews. There is an arms race between computational fake review detection and generation. In 2011, simple linguistic features such as n-grams and LIWC features [Pennebaker et al., 2001] were sufficient to detect fake reviews generated by humans [Ott et al., 2011]. In 2017, with advancements in deep learning, AI can generate fake Yelp reviews that are “not only evade human detection, but also score high on ‘usefulness’ metrics by users” [Yao et al., 2017b]. Finally, reputation systems suffer from the cold start problem, which refers to the fact that there are no reviews when systems initialize [Tavakolifard and Almeroth, 2012].

2.3.2 Networked Trust

Although online trust is not a totally new phenomenon, new challenges continue to arise as digitalized exchange platforms evolve. There is a narrative that there has been an decline in trust in the past few decades, especially trust in the government [Pew, 2019, Miller, 1974, Chanley et al., 2000]. According to a 2019 survey by Pew Research Center, public trust in the government remains near historic lows. “Only 17% of Americans today say they can trust the government in Washington to do what is right ‘just about always’ (3%) or ‘most of the time’ (14%).” [Pew, 2019]. However, at the same time, people engage in social exchange that require high levels of trust at scale — interactions that were totally unimaginable just a few years back (e.g., staying at strangers’ homes in the case of Airbnb). At the same time, even though traditional community-based organizations have seen decline in memberships (e.g., bowling leagues [Putnam, 1995]), online communities have exploded [Kraut and Resnick, 2012, Adamic et al., 2008, Zhang et al., 2007, Lampe et al., 2006]. Therefore, it is not informative to simply discuss whether there has been a “decrease” or “increase” in trust in the society, but rather, how trust structurally shifts from centralized institutions towards trust in networked environments.

In my view, fundamental changes to trust, driven by the digitalization of exchange, took place in three waves. (1) the digitization of the exchange of goods (e.g., Amazon, eBay); (2) the digitization of social relationships (e.g., Facebook, Twitter, LinkedIn, Instagram); and (3) the digitalization of the exchange of resources through sharing economy platforms (e.g., Airbnb, Uber, Lyft).

In the first wave, people have to trust each other primarily with their *financial resources*. Mechanisms to reduce risks, such as third-party payment systems and

reputation systems have overall successfully addressed the trust problem. New questions are still being raised, though, as online commerce platforms move towards more mobile-based where user-generated content play a more important role in establishing trust.

In the second wave, people have to trust others with their *personal information*. The social nature of these online platforms, primarily social network sites (SNSs), means that profiles play a central role in establishing trust. As people rely on on-line profiles rather than signals that are available in face-to-face communications, there are a lot of questions about how cues in Computer-Mediated Communication (CMC) establish trust. At the same time, through SNSs, people are able to maintain more weak ties, leading to “networked individualism”. We shift away from small, isolated, and tightly-knit groups to networked individuals [Rainie and Wellman, 2012, Granovetter, 1973]. At the same time, because SNSs make social relationships visible and data available, network analysis can reveal how network structures regulate information cascade [Cheng et al., 2014] and predict social and economic outcomes [Easley et al., 2010, Backstrom et al., 2006].

In the third wave, people have to trust others not only with their financial resources, personal information, but also *physical safety*. The wide adoption of sharing economy platforms is a relatively recent phenomenon, and they blur the line of social and economic exchange [Hamari et al., 2016, Ikkala and Lampinen, 2015, Lampinen and Cheshire, 2016]. Both profiles and reputation systems play an important role in establishing trust in sharing economy. Compared to SNSs, sharing economy platforms may involve higher risks (physical safety). At the same time, people tap into the benefits of conducting exchange with strangers on sharing economy platforms, rather than with their social ties, weak or strong.

Therefore, self-disclosure may be even more important to establish initial trust that is required for interactions on sharing economy platforms.

In addition to these three waves of the digitalization of exchange, another factor is playing an increasingly important role in mediating social exchanges — **algorithms**. Online platforms deploy algorithms through personalization and recommendation systems to optimize for specific engagement goals, and these algorithms affect how people interact and trust each other. For example, algorithms can match people in online dating sites [Hutson et al., 2018], curate what people see and impact how they make sense of social relationships on Facebook news feed [Eslami et al., 2015], and select to hide or display specific Yelp reviews [Eslami et al., 2019]. Algorithmic mediation creates more uncertainty in social exchange, potentially hindering interpersonal trust.

In short, the three waves of digitalization of exchange and the increasing algorithmic mediation create new questions about trust and call for a new way of thinking and reasoning about trust going forward.

Networked trust is such framework for thinking about trust online going forward. Borrowing from the term of “networked individualism” that describes the drastic shift in social structure brought by digital platforms [Rainie and Wellman, 2012], “**networked trust**” describes interpersonal trust in digitalized social exchange where cues, networks, and algorithms play important roles. Networked trust has three focuses: (1) *cues* in Computer-Mediated Communication; (2) embeddedness in social *networks*; and (3) increasing mediation by *algorithms*.

First, **cues in Computer-Mediated Communication** refers to interpersonal

trust where the initial social exchange takes place online and not face-to-face, following the rules of Computer-Mediated Communication. This aspect of networked trust is relatively well established, as I reviewed before in Section 2.3.1 on online trust. Two chapters of this dissertation, Chapter 3 and Chapter 4 expand on the focus on cues in networked trust — analyzing image and language cues at scale in commerce and sharing economy platforms respectively.

The second focus of networked trust, **embeddedness in social networks**, refers to the fact that our social exchange is embedded in social networks. For example, on LinkedIn, when users view another person’s profile, LinkedIn can reveal to users information of the network, such as how two people are connected or who their common connections are. Chapter 5 expands on the networks focus of networked trust, using the context of Facebook groups.

Finally, the third focus of networked trust, **increasing mediation by algorithms**, refers to the increasing control by algorithms on how and with whom we conduct exchange with. This focus is still a nascent area of research, closely related to the ongoing research on the issues of algorithmic fairness, accountability, and transparency (FAT*). In Chapter 6, I discuss future work that focuses on understanding how algorithms mediate exchange and impact interpersonal trust.

2.4 Bias and Distrust

In the last section of the chapter, I review literature on the flip side of trust — bias and distrust. As the process of forming trust is usually subjective, bias is closely related to trust. This section reviews work demonstrating four different

types of biases, including racial, gender, status, and algorithmic bias. Given the vast amount of literature on bias, I do not attempt to provide a comprehensive review but rather focus on studies on marketplaces or closely related to trust.

Racial bias. Research has observed racial bias in labor markets through a field experiment manipulating names on resumes [Lavergne and Mullainathan, 2004]. In the study, for resumes with similar credentials, white sounding names received 50 percent more callbacks for interviews than African-American sounding names across occupation, industry, and employer size. Racial bias is also observed on Airbnb through a field experiment by manipulating names in online profiles [Edelman et al., 2017]. Booking requests from guests with distinctively African-American names are roughly 16% less likely to be accepted than identical guests with distinctively white names.

In addition to names, visual information can also lead to racial bias online. One experiment showed that photos in online marketplaces with a dark-skinned hand holding the product lead to fewer and lower offers than photos with a light-skinned hand, indicating lower trust [Doleac and Stein, 2013]. Finally, in freelance marketplaces such as TaskRabbit, black workers receive significantly fewer reviews and worse ratings [Hannak et al., 2017].

Gender bias. One of the most prominent gender bias in marketplaces is the gender pay gap in labor markets. Women roughly earn 80 cents on average for each dollar earned by men in the labor market [Blau and Kahn, 2007, Dubey et al., 2017, Chamberlain, 2016]. Women sellers on eBay also receive less bids and lower final prices in auctions compared to men counterparts [Kricheli-Katz and Regev, 2016].

Representation and participation bias are also present in gender. In 2012, Vasilescu et al. provided empirical evidence showing that women on StackOverflow represent the minority of contributors and they participated less and earn less reputation [Vasilescu et al., 2012]. Similar gender representation gap was reported in editors of Wikipedia — a survey in 2010 found that fewer than 13% of Wikipedia contributors were women [Antin et al., 2011].

Status bias. Bias with regard to socio-economic status has been observed on sharing economy platforms. Research found that sharing economy is significantly more effective in dense, high socio-economic status areas than in low-socio-economic status areas [Thebault-Spieker et al., 2017]. For example, lower socio-economic status areas have higher wait times on Uber; and workers on TaskRabbit are less willing to travel to low socio-economic status areas for work and charge higher prices in those neighborhoods [Thebault-Spieker et al., 2017].

Status bias can also be context-induced. Contextual status can be induced by power imbalance in specific situations. For example, in the context of Couchsurfing, hosts have higher status compared to guests because the hosts provide free lodging for the guests. As a result, contextual status differences arise, and hosts receive higher ratings than guests in the same interactions, potentially due to the mechanism of status-giving in power dependence theory [State et al., 2016].

Algorithmic bias. The last type of bias has only recently begun to be better understood — algorithmic bias. In the 2016 paper, *Big Data's Disparate Impact*, Barocas and Selbst laid out how data mining can exhibit discrimination that can be unintentional and hard to understand. As a result, current law and techniques are inadequate in identifying and regulating the problem of algorithmic

bias [Barocas and Selbst, 2016].

Recent literature uncovers subtle racial and gender biases exhibited by algorithms [Buolamwini and Gebru, 2018, Datta et al., 2018, Bolukbasi et al., 2016, Eckhouse et al., 2019]. For example, an audit of two standard benchmark facial recognition datasets and three commercial systems revealed that the datasets are overwhelmingly composed of lighter-skinned subjects and the commercial systems have much higher rate of error for darker-skinned females than lighter-skinned men (34.7% v.s. 0.8%) [Buolamwini and Gebru, 2018]. Bolukbasi et al. also showed that word embeddings, a popular technique in natural language processing, could exhibit gender stereotypes such as associating “receptionist” to “female” [Bolukbasi et al., 2016].

Algorithmic bias raises the question: how might algorithmic bias affect trust? Work in transparency and interpretability of machine learning examines bias and trust in algorithms [Miller, 2018, Kizilcec, 2016, Eslami et al., 2015, Eslami et al., 2019]. Another line of research is policy-driven [Levy and Barocas, 2017, Hutson et al., 2018], outlining design and policy choices for platforms to make themselves less conducive to discrimination by users to gain user trust. At the same time, algorithmic bias might hinder interpersonal trust when the algorithms are crucial in mediating social exchange relationships.

2.5 Summary

Gambetta once wrote:

“The importance of trust pervades the most diverse situations where

cooperation is at one and the same time a vital and a fragile commodity: from marriage to economic development, from buying a second-hand car to international affairs, from the minutiae of social life to the continuation of life on earth." [Gambetta, 1988]

In this chapter, I reviewed prior work across many disciplines on trust in "diverse situations". By anchoring the definition of trust in social exchange theory, the work presented in the rest of the dissertation focuses on addressing new questions that the digitalization of exchange brings. Literature on different levels of trust was reviewed, including trust at the individual level, trust in groups, trust in platforms, trust in organizations, and trust at societal level. Finally, the framework of "networked trust" was proposed, which has three focuses: (1) *cues* in Computer-Mediated Communication; (2) embeddedness in social *networks*; and (3) increasing mediation by *algorithms*. As trust is a subjective process, we need to caution against biases, especially the ways that racial, gender, status, and socio-economic biases can manifest subtly in algorithms that mediate social exchange relationships.

CHAPTER 3

IMAGES OF TRUST: UNDERSTANDING IMAGE QUALITY AND TRUST IN PEER-TO-PEER MARKETPLACES

3.1 Introduction

In this chapter and the chapter that follows, I present two case studies that focus on the first focus of networked trust: trust in interactions where cues in Computer-Mediated Communication play an important role.

Specifically, this chapter explores how **image cues** establish trust in the context of second-hand peer-to-peer **online marketplaces**. As mentioned in Chapter 2, online marketplaces represent the first wave of the digitalization of social exchange, the digitization of the exchange of goods. Trust in the context of online marketplaces was the earliest focus of the discussion on online trust. Reputation systems were pioneered first in online marketplaces to facilitate online trust and research has discussed their benefits and limitations [Resnick et al., 2000, Resnick and Zeckhauser, 2002, Resnick et al., 2006, Cook et al., 2009].

However, as online commerce platforms continue to evolve and new technologies emerge, new challenges and opportunities about understanding trust on online commerce platforms arise. Increasingly, these platforms are becoming mobile-based, where user-generated images play an ever increasingly important role, compared to stock imagery. Although earlier work has studied the role of images in online marketplaces computationally [Goswami et al., 2011, Chung et al., 2012], user-generated images are less well understood. At the same time, advancements in computer vision (e.g., deep learning) present new opportunities

to understand image cues with more advanced techniques.

This chapter leverages such opportunities brought by advancements in computer vision to present a study on how user-generated images establish trust in second-hand peer-to-peer marketplaces.¹ Trust here is discussed in the context of trust in platforms, and was measured with a survey. The survey asked about the perceived trustworthiness in hypothetical marketplaces where images vary.

At its core, images are representations of the products to be exchanged online. Since it is impractical for a customer to inspect an item in person before purchasing it, images are very useful to establish trust by setting expectations, reducing uncertainty, and limiting information asymmetry in online marketplaces [Akerlof, 1978]. “Good” product images achieve the goals of expectation setting and uncertainty reduction effectively, and are likely to be successful in facilitating exchange.

Traditionally, stock images are considered as the gold standard for “good” images in online marketplaces, which are usually shot in professional studios with high-end equipment. They present the products in the best possible light, so to speak.

At the same time, as mobile-based peer-to-peer marketplaces such as LetGo.com and Facebook Marketplace gain popularity, the definition of “good” images may have evolved. These mobile-based applications enable easy listing process: users can quickly snap a picture with their phone, type in a short description, and instantly post their offerings. The peer-to-peer nature of these marketplaces also mean that sellers are often non-professionals, who either lack

¹This work was published at WACV 2019 as *Understanding Image Quality and Trust in Peer-to-Peer Marketplaces* [Ma et al., 2019b].

the expertise or motivation to provide a stock image style picture. As a result, peer-to-peer marketplaces today often contain images of mixed-quality that are user-generated rather than stock images.

In this context, it is important to understand how mixed-quality user-generated images establish trust. Specifically, I ask the following questions: Is it possible to have a definition of “good” user-generated product images that human raters agree on reliability? In addition, is it possible to computationally tell “good” and “bad” images apart accurately? And finally, how does image quality affect market outcomes, especially trust and sales? The rest of the chapter addresses these questions.

Specifically, this chapter makes the following contributions.

Annotating image quality. We curated a dataset of user-generated images of two common second-hand products ($\approx 75,000$ images). Then an annotating process was iteratively developed and we show that human raters can provide reliable judgements on on image quality under this process. We further annotated a third of the collected data with human-labeled quality judgment.

Modeling and predicting image quality. Previous work studied general photo aesthetics, but we show existing models do not completely capture image quality in the context of online marketplaces. In this work, we used both black-box neural networks and interpretable regression techniques to model and predict image quality (how appealing the product image appears to customers). For the regression-based approach, we model the factors of the photographic environment that influence quality with handcrafted features, which can guide potential sellers to take better pictures. As a result, we developed a better



Figure 3.1: We study the interplay between image quality, marketplace outcomes, and user trust in peer-to-peer online marketplaces. Here, we show how image quality (as measured by a deep-learned CNN model) correlates with user trust. User studies in Sec. 3.6.2 show that high quality images selected by our model out-performs stock-imagery in eliciting user trust.

understanding of how visual features impact image quality, while training a convolutional network to predict image quality with decent accuracy ($\approx 87\%$).

Marketplace outcomes: sales and perceived trustworthiness. Using our learned quality model, we then showed that image quality scores are associated with two different group outcomes: sales and perceived trustworthiness. Predicted image quality was associated with higher likelihood that an item is sold, while high quality user-generated images selected by our model out-performs



Figure 3.2: Samples of lowest-rated and highest-rated images from shoes and handbags groundtruth.

stock imagery in eliciting perceived trust from users (see Figure 3.1). Our findings can be valuable for designing better online marketplaces, or to help sellers with less photographic experience take better product images.

3.2 Related Work

Image quality in online marketplaces. Previous work has shown that image-based features such as brightness, contrast, and background lightness contribute to the prediction of click-through rate (CTR) in the context of product search [Goswami et al., 2011, Chung et al., 2012]. This line of work used hand-crafted image features, but did not actually assess the image quality as dependent variable. In another work on eCommerce, image quality was modelled and predicted through linear regression, and shown to be significant predictors of buyer interest [Di et al., 2014]. However, the dataset is not available, nor any details on the modelling methodology or model performance. Our work fills this gap by contributing a large annotated image quality dataset, along with improved model performance.

Automatic assessment of aesthetics. Another line of closely related work is

automatically assessing the aesthetic quality of images. Early work on aesthetic quality frequently used handcrafted features [Datta et al., 2006, Li et al., 2010, Bhattacharya et al., 2010]. More recent work focused on AVA (A Large-Scale Database for Aesthetic Visual Analysis), a large-scale dataset containing 250,000 images with aesthetic ratings [Murray et al., 2012] and adapting deep learning features [Lu et al., 2014] to improve prediction accuracy. In addition, aesthetics quality has been shown to predict the performance of images in the context of photography, increasing the likelihood that a picture will receive “likes” or “favorites” [Schwarz et al., 2016]. Although aesthetic quality is a fundamental and important feature of imagery, images online marketplaces belong to another visual domain compared to most of the images in this line of work. We show that aesthetic quality models do not completely capture product image quality.

Web aesthetics and trust. A large body of work in human computer interaction has investigated the link between the marketplace website’s aesthetics and perceived credibility. For example, previous work has shown low level statistics such as balance and symmetry correlate with aesthetics [Zheng et al., 2009, Michailidou et al., 2008]. More recently, computational models have been developed to capture visual complexity and colorfulness of website screenshots to predict aesthetic appeal [Reinecke et al., 2013]. However, since product images take up majority of the space in online marketplaces, inadequate attention has been paid to how the image quality, rather than interface quality, impact user trust. Our work focuses on product images as the most salient visual element of online marketplaces to study how they contribute to user trust.

Our work is concerned with *user trust*. Perceptions of user trust is important for several reasons: trust influences loyalty, purchase intentions [Hong and Cho,

2011], retention [Sun, 2010], and is important for the platform’s initial adoption and growth [Choi and Mai, 2018]. This is why we take a “social science” approach when soliciting trust judgments in Sec. 3.6.2.

Domain adaption: matching street photos to stock photos. One might wonder why marketplaces do not simply retrieve the stock photo of the product being depicted and display it instead. There are some problems with this approach. First, in used goods markets, stock photos do not depict the actual item being sold and are generally discouraged [Bland et al., 2007]. Second, stock image retrieval is a computationally challenging task, with state-of-the-art methods [Wang et al., 2016] achieving around 40% top-20 retrieval accuracy on the Street2Shop [Kiapour et al., 2015] dataset. Our work also contributes to the dataset of “street” photos taken by different users using a variety of mobile devices in varying lighting conditions and backgrounds.

3.3 Datasets

As shown in previous work [Di et al., 2014], image quality matters more for product categories that are inherently more visual (e.g., clothing). Thus, in our development of the dataset, we focus on the *shoes* and *handbags* categories. These two categories are among the most popular goods found on secondhand marketplaces and are visually distinctive enough to pose an interesting computer vision challenge.

There are two sources for the data used in this work: LetGo.com, and eBay. We focused on the publicly available LetGo.com images for creating the hand-annotation dataset, and used private data from eBay to test the relationship

between image quality and marketplace outcome — sales.

3.3.1 LetGo.com

LetGo.com is a mobile-based peer-to-peer marketplace for used items, similar to Craigslist. Potential buyers can browse through the “listings” made by sellers and contact the seller to complete the transaction out of the platform.

We collected product images data for two product categories, shoes and handbags. We crawled the front page of LetGo.com every ten to thirty minutes for a month, filtering the listings by relevant keywords in the product listing caption. For *shoes*, we used the keywords “shoe,” “sandal,” “boot,” or “sneaker” and collected data between November to December 2017 (66,752 listings containing 133,783 images in total). For *handbags*, we used the keywords “purse” or “handbag” and collected data between April to May 2018 (29,839 listings containing 44,725 images in total).

3.3.2 eBay

To understand whether image quality impacts real world outcomes, we partnered with eBay, one of the largest online marketplaces.

We collected data for listings on eBay in our two product categories, shoes and handbags, including the product images, meta-data associated with the listing, as well as whether the listing had at least one sales completed before becoming expired. We sampled data based on the date on which the listing expired (during

May 2018). We also down-sampled the available listings to create a balanced set of sold and unsold listings. In summary, this dataset included 66,000 sold and 66,000 unsold listings for *shoes*, and 16,000 sold and 16,000 unsold listings for *handbags*.

To evaluate our model’s generalizability in a real-world setting, we train models on publicly available LetGo.com data and test the relationship between predicted image quality and sales on eBay.

3.4 Annotating Image Quality

We collected ground truth image quality labels for our LetGo.com dataset (LetGo below) using a crowdsourcing approach. We designed the following task and issued it on Amazon Mechanical Turk, paying \$0.8 per task. Each task contained 50 LetGo images randomly batched, and at least 3 workers rated each image. For each image, the worker was asked to rate the image quality on a scale between 1 (not appealing) to 5 (appealing). In total, we annotated 12,515 images from the *shoes* category and 12,222 from the *handbags* category. Only LetGo data was used for this task.

An important consideration is the difference between *product* quality and *photographic* quality. In this survey, we are primarily interested in what the merchant can do to make their listings more appealing, so it is important that workers ignore perceived product differences. To help prime our workers along this line of thinking, we added two text survey questions spaced throughout each task with the following prompt: “Suppose your friend is using this photo to sell their shoes. What advice would you give them to make it a better picture? How can

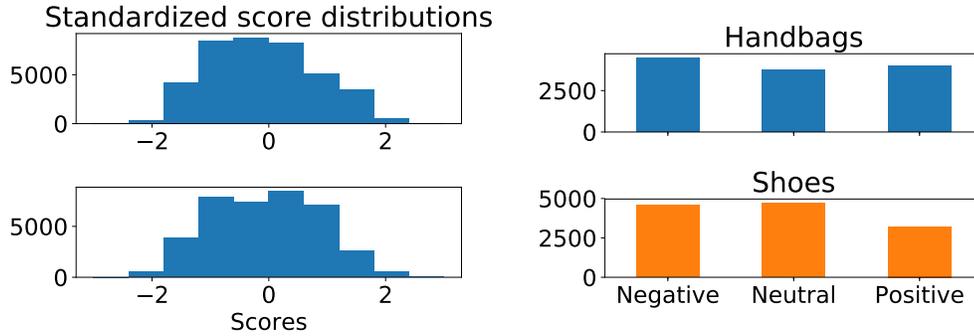


Figure 3.3: Left: standardized score distributions for filtered images on shoes and handbags categories. Right: final ground truth labels, showing a fairly even distribution.

the seller improve this photo?" This task also slows workers down and forces them to carefully consider their choices.

After collecting the data, we first standardized each worker’s score distribution to zero mean and unit variance to account for task subjectivity and individual rater preference. Then, we filtered out images where the standard deviation of all rater scores was within the top 40%, retaining 60% of the original dataset. This removed images where annotators strongly disagree. This filtering also improved the inter-rater agreement as measured by an average pairwise Pearson’s ρ correlation across each rater from 0.34 ± 0.0046 on the unfiltered shoes data to 0.70 ± 0.0031 on the filtered shoes data. This shows our labeling mechanism can reliably filter images with low annotation agreement.

Finally, we discretized the scores into three image quality labels: good, neutral and bad. To do this, we rounded the average score to the nearest integer, and took the positive ones as good, zeros as neutral, and negative ones as bad. The resulting *shoes* and *handbags* datasets are roughly balanced; see score distributions in Figure 3.3 and example images in Figure 3.2.

3.5 Modeling Image Quality

After collecting ground truth for our dataset, we study what factors of an image can influence perceived quality. This turns out to be nontrivial. Shopping behavior is complicated, and customers and crowd workers alike may have intricate preferences, behaviors, and constraints.

We attempt to model image quality in two ways: first, we train a deep-learned CNN to predict ground truth labels. This model is fairly accurate, allowing us to approximate image quality on the rest of our dataset, but as a “black box” model it is largely uninterpretable, meaning it does not reveal what high-level image factors lead to high image quality. Second, to understand quality at an interpretable level, we then use multiple ordered logistic regression to predict the dense quality scores. This lets us draw conclusions about what photographic aspects lead to perceived image quality.

Prediction. For our model, we use the pretrained *Inception v3* network architecture [Szegedy et al., 2016] provided by PyTorch [Paszke et al., 2017]. Our task is 3-way classification: given a product image, the model predicts one of $\{0, 1, 2\}$ for bad, neutral, and good respectively. To do this, we remove the last fully-connected layer and replace it with a linear map down to 3 output dimensions.

Image quality measurements are subjective. Even though our crowdsourcing worker normalization filters data where workers disagree, we want to allow the model to learn some notion of uncertainty. To do this, we train using *label smoothing*, described in § 7.5 of [Goodfellow et al., 2016]. We modify the negative log likelihood loss as follows. Given a single input image, let x_i for $i \in \{0, 1, 2\}$ be

the three raw output scores (logits) from the model and let $\hat{x}_i = \log \exp x_i / \sum_j \exp x_j$ be the output log-probabilities after softmax. To predict class $c \in \{0, 1, 2\}$, we use a modified label *smoothing loss*, $\ell(x, c) = -(1 - \varepsilon)\hat{x}_c - \varepsilon \sum_{i \neq c} \hat{x}_i$ for some smoothing parameter ε , usually set to 0.05 in our experiments. This modified loss function avoids penalizing the network too much for incorrect classifications that are overly confident.

These models were fine-tuned on our LetGo dataset for 20 epochs (*shoes*: $N=12,515$; *handbags*: $N=12,222$). We opted not to use a learning rate schedule due to the small amount of data.

Aesthetic quality baseline. We also considered a baseline aesthetic quality prediction task to test whether existing models that capture photographic aesthetics can generalize to predict product image quality. We fine-tuned an Inception v3 network on the AVA Dataset [Murray et al., 2012]. To transform predictions into outputs suited to our dataset, we binned the mean aesthetic annotation score from AVA into positive, neutral, and negative labels. The model was fine-tuned on AVA for 5 epochs.

Evaluation. We used a binary “forced-choice” accuracy metric: starting from a held-out set of positive and negative examples, we considered the network output to be positive if $x_2 > x_0$, effectively forcing the network to decide whether the positive elements outweigh the negative ones, removing the neutral output. By this metric, our best shoes model achieved 84.34% accuracy and our best handbag model achieved 89.53%. This indicates that our model has a reasonable chance of agreeing with crowd workers about the overall quality sentiment of a listing image. (If we include neutral images and predictions and simply compare the 3-way predicted output to the ground truth label on the entire evaluation

set, our handbag model achieves 64.09% top-1 accuracy and our shoes model achieves 58.36%.) For comparison, the baseline aesthetic model achieved 68.8% accuracy on shoes images and 78.8% accuracy on handbags. This shows that product image quality cannot necessarily be predicted by aesthetic judgments alone, and that our dataset constitutes a unique source of data for the study of online marketplace images. Our image quality model gives us a black box understanding of image quality in an uninterpretable way. Next, we investigate what factors influence image quality.

3.5.1 What Makes a Good Product Photo?

Guiding sellers to upload good photos for their listings is a challenge that almost all online eCommerce sites face. Many sites provide photo tips or guidelines for sellers. For example, Google Merchant Center [Google, 2019]. suggests to “use a solid white or transparent background”, or to “use an image that shows a clear view of the main product being sold”. eBay also provides “tips for taking photos that sell” [eBay, 2019], including “tip #1: use a plain, uncluttered backdrop to make your items stand out”, “tip #2: turn off the flash and use soft, diffused lighting”. In addition, many online blogs and YouTube channels also provide tutorials on how to take better product photos.

Despite the abundance of product photography tips, little previous work has validated the effectiveness of these strategies computationally (with the exception of [Chung et al., 2012, Goswami et al., 2011]). Although there is a robust line of research on computational photo aesthetics (e.g., [Schwarz et al., 2016]), product photography differs greatly in content and functionality from

other types of photography and is worth special examination.

In this work, we leverage our annotated dataset, and conducted the first computational analysis of the impact of common product photography tips on image quality. Unlike previous work [Chung et al., 2012, Goswami et al., 2011] that analyzed the impact of image features on clicks, we evaluate directly on potential buyers' perception of image quality. In a later section, we then show how image quality can in turn predict sales outcomes.

Selecting Image Features

In order to select the image features to validate, we took a qualitative approach and analyzed 49 online product photography tutorials. We collected the tutorials through Google search queries such as "product photography", "how to take shoe photo sell", and took results from top two pages (filtering out ads). We manually read and labeled the topics mentioned in these tutorials and summarized the most frequently mentioned tips. Out of the 49 tutorials we analyzed, the most frequent topics were: (1) Background (mentioned in 57% of the tutorials): keywords included white, clean, uncluttered; (2) Lighting (57%): soft, good, bright; (3) Angles (40%): multiple angles, front, back, top, bottom, details; (4) Context (29%): in use; (5) Focus (22%): sharp, high resolution; (6) Post-production (22%): white balance, lighting, exposure; and (7) Crop (14%): zoom, scale.

Based on the qualitative results, as well as referencing previous work [Chung et al., 2012, Goswami et al., 2011], we defined and calculated a set of image features to analyze for their impact on image quality. There are three types of image features that we considered: (1) Global features such as brightness,

contrast, and dynamic range; (2) Object features based on our object detector (more details on that below); and (3) Regional features focusing on background and foreground. Table 3.1 contains the definition and example images for our complete set of image features.

Calculating Image Features

Global image features can be computed without extra information, but object and regional features require knowledge of the object that appears in the image. We trained an object detector that could detect bounding boxes for our product categories.

Object detection. The process for building shoe detectors and handbag detectors is the same. First, we collected and manually verified a dataset of 170 shoe images from the ImageNet dataset [Deng et al., 2009]. Those images were already labeled with the bounding box around each shoe, and they vary across many visual styles and contexts (not just online marketplaces). We also augmented the ImageNet images with our own from the online marketplace mentioned in Sec. 3.3.2. We designed a crowdsourcing task and labeled 500 images from the online marketplaces image dataset. Crowd workers were asked to draw the bounding box around each single shoe in the image, and each image was assigned to two distinct crowd-workers in order to ensure quality labeling. We filtered out labels where the overlap between two bounding boxes were less than 50%. In total, we gathered 650 shoe images from both ImageNet and our online marketplace datasets, with bounding boxes around each shoe.

Next, we trained our shoe detector by using the Tensorflow Object Detection

| Feature Name | Definition | Low | High |
|--|--|---|---|
| Global Features: | | | |
| brightness | $0.3R + 0.6G + 0.1B$ |  |  |
| contrast | Michelson contrast |  |  |
| dynamic_range | grayscale (max - min) |  |  |
| width | the width of the photo in px | | |
| height | the height of the photo in px | | |
| resolution | $\text{width} * \text{height} / 10^6$ | | |
| Object Features: | | | |
| object_cnt | # of objects detected |  |  |
| top_space | bounding box top to top of image in px |  |  |
| bottom_space | bounding box bottom to bottom of image |  |  |
| left_space | bounding box left to left of image |  |  |
| right_space | bounding box right to right of image |  |  |
| x_asymmetry | $\text{abs}(\text{right_space} - \text{left_space}) / \text{width}$ | | |
| y_asymmetry | $\text{abs}(\text{top_space} - \text{bottom_space}) / \text{height}$ | | |
| Regional Features: (fg: foreground; bg: background) | | | |
| fgbg_area_ratio | # pixels in fg / bg |  |  |
| bgfg_brightness_diff | brightness of bg - fg |  |  |
| bgfg_contrast_diff | contrast of bg - fg |  |  |
| bg_lightness | RGB distance from a pure white image |  |  |
| bg_nonuniformity | standard deviation of bg pixels in grayscale |  |  |

Table 3.1: Image feature definitions and example images

API and Single Shot Multi-box detector method [Liu et al., 2016]. The training set included 300 images randomly selected from our labeled datasets. Finally, we evaluated the performance of our detector on a validation set of 350 images. We achieved a mean Average Precision (mAP0.5) of 0.84 for the shoe detection.² We repeated the same process for the handbag detector, and reached similar performance.

The resulted object detectors output bounding boxes around the shoes and handbags in the image, which we then used to compute object and regional features. In particular, for regional features, we used GrabCut algorithm [Rother et al., 2004] to segment the foreground and background, initializing GrabCut with the detected bounding boxes as the foreground region. Then we computed the lightness and non-uniformity of the background, as well as differences in brightness and contrast between background and foreground. Table 3.1 contains all details and example images for our image features.

Regression Analysis

After calculating the image features, we analyzed their impact on image quality through multiple ordered logistic regression. Our dependent variable is the three image quality labels (bad, neutral, good) annotated through the crowdsourcing task in Sec. 3.4. We choose to use the image label rather than raw image quality scores for the dependent variable because there is less noise in the label (as we did majority voting to get image labels). All analysis was done on the LetGo image dataset, as that's the set of images we collected manual image quality label on.

²mAP0.5 means an image counts as positive if the detected and groundtruth bounding boxes overlap with an intersection-over-union score greater than 0.5.

Detection: shoes. We first report on analysis of the shoe images. In our dataset, 7% of the images did not have a target object detected, 22% has one target object detected, 70% had two or more targeted objects detected. A chi-squared test showed that there were significant differences in the distribution of image quality labels across different number of target objects detected ($\chi^2 = 463.68$, $p < .001$). Having at least one target object detected makes the image 2.7 times more likely to be labeled as “good” quality.

Detection: handbags. Similarly, for handbags, 8% of our images did not have a target object detected (with the detection score threshold of .90). A chi-squared test showed that there was a significant difference in the distribution of image quality labels between having a target object detected and not ($\chi^2 = 60.34$, $p < .001$). Having the target object detected makes the image 1.4 times more likely to be labeled as “good” quality.

To ensure the regression analysis remained accurate, we manually verified detection accuracy on a subset of 2,000 images of shoes and 2,000 images of handbags. We then conducted ordered logistic regression on this manually-verified subset using image features as independent variables to predict the image quality label. Results are shown in Table 3.2³.

Brightness/background. On a high level, we confirmed brighter images are more likely to be labeled as high quality for both product categories. In addition, the non-uniformity of the background makes it less likely for an image to be labeled as high quality. These two features coincide with the most commonly mentioned product photography tips, background and lighting.

³Note the regression results do not differ significantly when we include the entire dataset, showing that features extracted from automatic object detection are robust.

| Feature Name | Shoes | | Handbags | |
|----------------------|----------|---------|----------|--------|
| | Estimate | SE | Estimate | SE |
| brightness | 3.30*** | (.96) | 3.46*** | (.92) |
| contrast | 1.79 | (1.26) | 4.89*** | (1.44) |
| dynamic_range | 2.22* | (1.104) | .47 | (1.67) |
| resolution | -.10 | (.06) | -.21 | (.19) |
| x_asymmetry | -0.54 | (.68) | -2.26** | (.84) |
| y_asymmetry | -0.74 | (.49) | .73 | (.46) |
| fgbg_area_ratio | -.35*** | (.06) | -.17*** | (.03) |
| bgfg_brightness_diff | .59 | (.48) | -.11 | (.53) |
| bgfg_contrast_diff | -.52 | (.54) | -.80 | (.42) |
| bg_lightness | -.46 | (.56) | .16 | (.55) |
| bg_nonuniformity | -1.6* | (.633) | -4.84*** | (.64) |
| 0 1 | 3.59** | (1.20) | 4.64*** | (1.22) |
| 1 2 | 5.46*** | (1.21) | 6.13*** | (1.22) |
| AIC: | 4170.76 | | 4169.92 | |

Significance codes: * $p < .05$, ** $p < .01$, *** $p < .001$

Table 3.2: Ordered logistic regression coefficients predicting image quality labels

Crop/zoom. We found mixed evidence around the crop of the images. For both product categories, higher foreground to background ratio makes it less likely for an image to be labeled as high quality, suggesting that the product should be properly framed and not too zoomed-in.

Symmetry. Interestingly, we found that for shoes, asymmetry does not significantly contribute to perception of quality, but for handbags, horizontal asymmetry moderately contributes to a lower perception of quality. One potential explanation for this difference could be that shoes and handbags have different product geometry dimensions (tall v.s. wide), and sellers would take pictures with different orientations resulting in different distribution in vertical and horizontal asymmetry to begin with. Indeed, we observed that handbag images were slightly more likely to be in portrait (width<height) orientation than landscape orientation (24% v.s. 22%, $\chi^2=28.6$, $p<.001$). Further investigation is necessary to

understand how symmetry impact the perceived quality of product photos.

Contrast. Finally, we observed that the difference in contrast between background and foreground can impact perceived image quality. In other words, a good quality product image’s background should have low contrast, and the foreground (the product) should have high contrast. The difference in brightness between background and foreground, or the lightness of the background were not significant in making an image more likely to be labeled as high quality — suggesting a uniform background, either dark or bright, could be both effective, and potentially work better for different colored products.

3.6 Marketplace Outcomes

In the previous section, we have shown that our trained models can automatically classify user-generated product photos with an accuracy of 87% (averaged across two product categories, shoes and handbags). Now leveraging the quality scores predicted using our models, we proceed to examine how image quality contributes to real-world and hypothetical marketplace outcomes.

We focus on two complementary marketplace outcomes: (1) *Sales*: whether an individual listing with higher quality photos is more likely to generate sales; and (2) *Perceived trustworthiness*: whether a marketplace with higher quality photos is perceived as more trustworthy.

The first outcome, sales, is naturally important for online marketplaces. As most of these platforms operate on a fee-based model (i.e., the platform charges a flat or percentage fee when an item is sold), sales are directly linked to the

revenue and success of the marketplaces. Further, whether an item is sold is also likely associated with higher user satisfaction.

Second, the perception of whether a platform is trustworthy is important for the platform's initial adoption and growth [Choi and Mai, 2018]. Previous work has shown that the perceived trustworthiness of online marketplaces has a strong influence on loyalty and purchase intentions of consumers [Hong and Cho, 2011], retention [Sun, 2010], as well as creating a price premium for the seller [Pavlou and Dimoka, 2006]. Several factors of the website's visual design are known to impact trust (e.g. complexity and colorfulness [Reinecke et al., 2013]), but the effect of product image quality on trustworthiness remains an open question.

3.6.1 Image Quality and Sales

For the first outcome (sales of individual listings), we used log data from eBay for analysis. Since merchants sometimes sell many quantities of the same item, we predict whether a listing sold at least once before it expired. To that end, a sample of balanced sold and unsold listings was created for the two product categories we studied — shoes and handbags — see details of data sampling in Sec. 3.3.2.

We first used logistic regression to understand how image quality relates to trust. We considered three different models: (1) a baseline model using metadata information about the listings, including the number of days the listing has been on the platform, listing view count, and the item price; (2) a model including image quality prediction score (the log-probability that an image will be classified as high quality by models trained on our dataset), in addition to baseline features;

| | Shoes ($N=130K$) | | | Handbags ($N=32K$) | | |
|----------------|--------------------|--------------------|--------------------|----------------------|--------------------|--------------------|
| | (1) | (2) | (3) | (1) | (2) | (3) |
| (Intercept) | -1.09*** (0.05) | -1.14*** (0.05) | -1.11*** (0.05) | -2.66*** (0.07) | -2.61*** (0.07) | -2.42*** (0.07) |
| # Days (Log) | 0.26*** (0.01) | 0.26*** (0.01) | 0.26*** (0.01) | 0.64*** (0.01) | 0.63*** (0.01) | 0.60*** (0.01) |
| # Views (Log) | 1.07*** (0.01) | 1.08*** (0.01) | 1.08*** (0.01) | 0.94*** (0.01) | 0.95*** (0.01) | 0.98*** (0.01) |
| Price (Log) | -0.57*** (0.01) | -0.56*** (0.01) | -0.57*** (0.01) | -0.31*** (0.01) | -0.32*** (0.01) | -0.36*** (0.01) |
| Img. Quality | | 0.16*** (0.01) | 0.15*** (0.01) | | 0.22*** (0.01) | 0.15*** (0.01) |
| Img. Aesthetic | | | 0.08*** (0.01) | | | 0.31*** (0.02) |
| AIC | 116,968 | 116,525 | 116,414 | 30,685 | 30,453 | 30,026 |

Note: * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

Table 3.3: Image quality predicted by our models is positively associated with higher likelihood that an item is sold (1.17x more for shoes, and 1.25x more for handbags).

and (3) a model including both image quality and aesthetic quality scores (trained on the AVA dataset), in addition to baseline features. The results of regressions for both product categories predicting whether an item is sold are reported in Table 3.3.

From the regression analysis, we show that image quality predicted by our models is associated with higher likelihood that an item is sold (odds ratio 1.17 for shoes, 1.25 for handbags, $p < .001$). These results hold even when controlling for the predicted aesthetic quality of images. Interestingly, both predicted image quality and aesthetic quality are more strongly associated with the sales of handbags than shoes ($\beta=0.22$ v.s. 0.16 for image quality, and $\beta=0.31$ v.s. 0.08 for handbags), potentially signaling that handbag is a more visual product category than shoe.

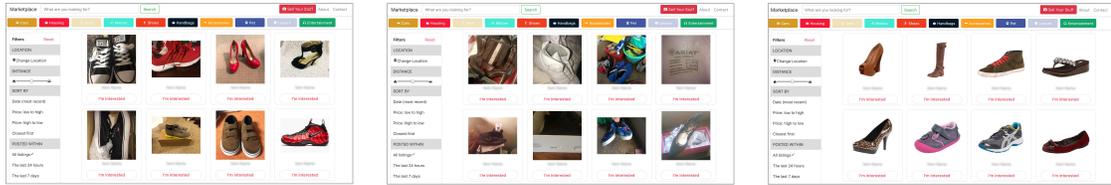


Figure 3.4: Hypothetical marketplace mock-ups used for our user experiment. From left to right, showing images with high quality score, low quality score, and stock imagery.

However, in terms of model performance, both image quality and aesthetic quality only resulted in small improvements to the baseline model. We illustrate the model performance through prediction accuracy — 10-fold cross validation showed that the baseline model can predict whether an item will be sold with an accuracy of 80.3% for shoes, and 74.7% for handbags. Both image quality and aesthetic quality improved the prediction accuracy only marginally (around 1%).

These findings suggest that both image quality and aesthetic quality are associated with higher likelihood of sales for individual listings on online marketplaces, though both have limited power in improving the accuracy of sales prediction. Future work could explore the difference among different product categories, or the relationship between image quality and other metrics for online marketplaces, such as “sellability” [Liu et al., 2017], and click-through-rate [Goswami et al., 2011, Chung et al., 2012].

3.6.2 Perceived Trustworthiness

For the second outcome, perceived trustworthiness of the marketplace, we designed a user experiment to compare the effects of good/bad quality user-

generated images compared to stock imagery. Our goal here is to show the potential application of our image quality models in improving the perceived trustworthiness of online marketplaces. Our hypothesis is that high quality marketplace images (as selected by our models) will lead to the highest perception of trust, followed by stock imagery, and then low quality marketplace images will lead to the lowest perception of trust. This hypothesis is rooted in the fact that uncertainty and information asymmetry are fundamental problems in online marketplaces that limit trust [Akerlof, 1978]. Stock images, though high in aesthetic quality, do not help reduce the uncertainty of the actual conditions of the product being sold. High quality user-generated images could help bridge the gap of information asymmetry and reduce uncertainty, therefore increasing the trust.

To test this hypothesis, we built three hypothetical marketplaces mimicking the features of popular online platforms (see Figure 3.4), each populated with (1) high quality marketplace images selected by our model, (2) low quality marketplace images selected by our model, and (3) stock imagery from the UT Zappos50K dataset [Yu and Grauman, 2014, Yu and Grauman, 2017]. Specifically, we designed a between-subject study with three conditions, varying the images of the listings (good; bad; and stock). Each participant was randomly assigned to one condition and saw three example mock-ups of a hypothetical online peer-to-peer marketplace website, each populated with 12 images randomly drawn from a set of 600 candidate images.

We prepared the candidate images for each condition in the following ways. All of our images were from the *shoes* category, but could easily be expanded to other categories. For the “good” condition, the candidate images were 600 images

randomly drawn from the top 5% LetGo images as predicted by our image quality model (to have high image quality). Similarly, candidate images used for the “bad” condition were drawn from the bottom 5% of LetGo images predicted by our image quality model. For the “stock” image condition, we randomly sampled 600 images from the UT Zappos50K [Yu and Grauman, 2014, Yu and Grauman, 2017] as candidate images.

The main dependent variable for this experiment is the perceived trustworthiness of the marketplace, which could be measured in a few different aspects. Therefore, we developed a six-item trust scale based on adaptations of previous work on trust in online marketplaces [Corbitt et al., 2003], shown in Table 3.4. Each participant was requested to rate the marketplace they were shown on a 5-point Likert scale. We take the average of participant’s responses to all items in the scale as the “trust in marketplace” score.

Results

The experiment was pre-approved by Cornell’s Institutional Review Board, under protocol #1805007979. We issued the task through Amazon Mechanical Turk and recruited 333 participants, paying 50 cents per task.

We retained 303 submissions after initial filtering, evenly distributed across three conditions. We filtered out the submissions that were completed too fast or too slow (trimming the top and bottom 5% based on task completion time), as previous work has shown that filtering based on completion velocity improves the quality of submissions [Ma et al., 2017b, Wilber et al., 2017].

Overall, participants reported highest level in marketplaces populated with

good images, followed by stock imagery, and the lowest level of trust in marketplaces populated with bad images ($p < .001$). The average perceived trustworthiness of marketplaces per condition is shown in Figure 3.1.

The “trust gap” between high quality user-generated images and stock imagery suggests that high “quality” images on online marketplaces do not necessarily have to be more aesthetically pleasing. Our finding corroborates previous findings that show users prefer actual images over stock imagery because they give an accurate depiction of what the product looks like [Bland et al., 2007]. Stock images could be perceived impersonal or “too good to be true” in an online peer-to-peer setting. High quality user-generated images, on the other hand, reduce uncertainty and information asymmetry in online transaction settings, therefore increasing trust.

Taken together, our marketplace experiments showed that our image quality models could effectively pick out images that are of high quality and can increase the perceived trustworthiness of online peer-to-peer marketplaces, even outperforming stock imagery. The results of the experiment also suggest potential real-world applications of our image quality dataset as well as prediction models, by automatically ranking, filtering and selecting high quality images to present to the users to elicit feelings of trust.

3.7 Discussion and Conclusion

This work presents a computational understanding of user-generated images in online marketplaces and how image quality relates to trust. By gathering and annotating a large-scale dataset of user-generated product images from

| Item | Definition |
|-------------|---|
| General | How trustworthy do you think this marketplace is? |
| Technical | I believe that the chance of having a technical failure on this marketplace is quite small. |
| Risk | I believe that online purchases from these sellers are risky. |
| Expectation | I believe that products from these sellers will meet my expectations when delivered. |
| Care | I believe that these sellers care about its customers. |
| Fidelity | I believe that the photos accurately represent the condition of the products. |

Note: Response to each item is based on a 5-point Likert scale (strongly disagree to strongly agree)

Table 3.4: Marketplace perceived trustworthiness scale

online marketplaces, we were able to develop a deeper understanding of the visual factors that improve image quality, while reaching a decent accuracy ($\approx 87\%$) for predicting image quality. In particular, high quality user-generated images selected by our model outperform stock imagery in eliciting perceived trustworthiness in the platform by a potential user.

These findings have important implications for the future of increasingly mobile-based online marketplaces; the dataset can also be useful for the broader computer vision community, e.g., providing more examples of real-world images for domain adaptation tasks. One can also leverage this work to build tools to help users take better product photos, potentially through novel interfaces such as Augmented Reality. These tools can potentially facilitate trust between sellers and buyers on online marketplaces.

The fact that high quality user-generated images outperform stock images highlights how the dynamics between images and trust have evolved over the years. User-generated images may continue to play a more important role, as users believe that they reduce more uncertainty and are more useful in expecta-

tion setting. At the same time, low quality user-generated images perform badly in generating perceived trustworthiness. Future platforms can consider algorithmically filtering out low-quality user-generated images, or provide suggestions on which images to feature for users to promote overall trust on the platform.

This work is also not without limitations. The dataset, while large in size, only covered two product categories. Predicted image quality also only has limited prediction power in whether an item is sold. Future work could further enrich the dataset, expanding to other categories, such as images of Airbnb listings. Future work can also explore how image quality might indirectly influence sales (e.g., through increased view count).

Finally, from a networked trust perspective, the work presented in this chapter furthers our understanding of the first aspect of networked trust (CMC). By comparing user-generated images with stock images, I show that trust online through images is not simply established with aesthetics or photographic quality, but rather, the principles of uncertainty reduction. While photographic quality definitely plays a role, and low-quality user-generated images perform badly in eliciting perceived trustworthiness, there might be a “uncanny valley” when it comes to image aesthetics and trust. When something appears “too good to be true”, such in the case of stock images, user trust might be negatively impacted. High-quality user-generated images, on the other hand, may have hit a sweet spot between aesthetic quality and expectation management and uncertainty reduction, thus gaining most trust.

Importantly, it is important to point out that it is a conscious choice to *not* focus on image cues of people but rather image cues of products in this research inquiry on the relationships between images and trust, as prior research has pointed out

racial bias in state-of-art facial recognition algorithms [Buolamwini and Gebru, 2018]. However, images of people are perhaps even more important in the context of interpersonal trust and trust in media and information. Future work can further investigate how image cues of people affect trust and its interaction with bias. For example, does images of faces influence how politicians are trusted? Is it possible to study such mechanisms computationally while accounting for racial and gender bias in current facial recognition systems? Finally, as computer vision techniques continue to advance, images and videos can both be manipulated to distort the true representation of people. For example, a doctored video of House Speaker Nancy Pelosi reached millions of viewers on Facebook [Rini, 2019]. Understanding how such manipulated image and video cues can impact interpersonal trust remains important future work for networked trust, especially in the context of misinformation.

CHAPTER 4
LANGUAGE OF TRUST: SELF-PRESENTATION AND PERCEIVED
TRUSTWORTHINESS OF AIRBNB HOST PROFILES

4.1 Introduction

The last chapter presented a study of trust in one of the more traditional types of online exchange platforms, online commerce platforms, where I showed that image cues are important to establish trust. In this chapter, we examine how **language cues** establish trust in a relatively new type of social exchange platform — **sharing economy**.

I view sharing economy platforms among the third wave of the digitalization of exchange — the digitalization of the exchange of resources. Challenges about trust in this wave of digitalization of exchange center around the fact that people need to have high levels of trust in *strangers*. Not only people need to trust strangers with their financial resources, personal information, but also *physical safety*. Language has been shown to be especially important in the early stages of relationship development among strangers, through the process of self-disclosure [Cozby, 1972]. As sharing economy platforms gain popularity, it is important to understand how language cues establish interpersonal trust in this context.

In particular, this chapter examines how language cues establish trust on Airbnb through self-disclosure in host profiles.¹ I discuss trust on Airbnb in a

¹The work presented in this chapter was published at CSCW 2017 as *Self-disclosure and Perceived Trustworthiness of Airbnb Host Profiles*, and in ICWSM 2017 as *A Computational Approach to Perceived Trustworthiness of Airbnb Host Profiles* [Ma et al., 2017a, Ma et al., 2017c].

dyadic context, and measure perceived trustworthiness of host using surveys.

Airbnb is an online lodging marketplace for short-term peer-to-peer rentals, facilitating monetary and social exchange between individuals [Lampinen and Cheshire, 2016]. On Airbnb, *hosts* can list places (e.g. rooms, apartments, houses, or even boats and castles) for *guests* to rent. The guest is often a temporary visitor, and is not acquainted with the host beyond Airbnb. As of July 2019, Airbnb reports six million listings, and 500 million guest arrivals all-time through the platform².

The main utility of Airbnb — identifying potential lodging resources offered by unknown individuals — comes with risks that affect both guests and hosts who wish to participate in the exchange. A potential host may worry about guests damaging their property. A potential guest may fret about their physical safety, the truthfulness of the quality of the property being advertised, or whether the host would be kind enough to provide assistance in exigencies [Ert et al., 2016]. Establishing guest-host trust helps manage such uncertainties and risks — making trust a crucial factor for the success of such social exchange sites.

There are several ways that Airbnb designs for trust. Airbnb has an assurance policy and a reputation system in place, in addition to making information about the host and property readily available before booking. On Airbnb, each host has a profile page that includes photos, a text-based self-description, social media verification status, and reviews (if any) from other Airbnb users who have stayed with the host. These profiles contribute to a guest’s decision-making process [Newman and Antin, 2016], and help establish perceived trustworthiness [Ert et al., 2016]. In this work, we focus on host profiles, especially the

²<https://press.airbnb.com/fast-facts/>

text-based self-description and its role in establishing the perceived trustworthiness of hosts in the eyes of potential guests.

There are two steps towards the understanding of how language cues establish trust on Airbnb. In the first step, we examine how hosts describe themselves on their Airbnb profile pages. We use a mixed-methods study to develop a categorization of the topics that hosts self-disclose in their profile descriptions, and show that these topics differ depending on the type of guest engagement expected. We also examine the perceived trustworthiness of profiles using topic-coded profiles from 1,200 hosts, showing that longer self-descriptions are perceived to be more trustworthy. Further, we show that there are common strategies (a mix of topics) hosts use in self-disclosure, and that these strategies cause differences in perceived trustworthiness scores. We then show that the perceived trustworthiness score is a significant predictor of host choice — especially for shorter profiles that show more variation. The results are consistent with uncertainty reduction theory, reflect on the assertions of signaling theory, and have important design implications for sharing economy platforms, especially those facilitating online-to-offline social exchange.

In the second step, we developed a novel computational framework to predict the perceived trustworthiness of host profile texts. To achieve this goal, we extended the previous dataset from 1,200 to 4,180 Airbnb host profiles annotated with perceived trustworthiness. To the best of our knowledge, the dataset along with our models allow for the first computational evaluation of perceived trustworthiness of textual profiles, which are ubiquitous in online peer-to-peer marketplaces. We provide insights into the linguistic factors that contribute to higher and lower perceived trustworthiness for profiles of different lengths.

4.2 Related Work

User profiles are an important part of many online systems, and serve a variety of functions [Uski and Lampinen, 2014]. In social networking sites, profiles provide an identity for the user that persists over time and the myriad of interactions on the site [boyd and Ellison, 2007]. In online dating sites, profiles provide self-disclosure that attracts the interest of other users, while limiting the risks associated with outright deception [Gibbs et al., 2010, Toma et al., 2008].

In each of these contexts — especially in services that lead to offline interactions — a key function of the profile is to establish perceived trustworthiness [Ert et al., 2016, Guttentag, 2015]. Here, we distinguish between *trustworthiness* and *trust*. Trustworthiness is an attribute of a trustee [Hardin, 2002, Kiyonari et al., 2006], while trust is exhibited by a trustor (e.g. a decision to take risks in an economic game [Berg et al., 1995, Cook et al., 2005]). As our focus is on the host, the trustee, we focus on perceived trustworthiness as an attribute of the host.

One approach to understanding how profiles are used to establish trustworthiness is the Profile as Promise framework [Ellison and Hancock, 2013], which uses the notion of a promise to characterize the function of a profile. In this view, the profile is a psychological contract between the profile holder, and the viewer that future interactions (e.g. with a date, a car driver, or an Airbnb host) will take place with someone who does not differ fundamentally from the person represented in the profile. The notion of a promise has been successfully applied to various contexts that require good faith to operate, including online dating [Ellison and Hancock, 2013] and job hunting [Rousseau and Greller, 1994].

The Profile as Promise perspective argues that the content and characteristics

of the disclosures, or promises, made in user profiles should be diagnostic of trustworthiness perceptions. Within this framework we draw on theories from communication, and from economics, to form specific expectations about disclosures and their perceptions. For example, communication scholars have used uncertainty reduction theory [Berger and Calabrese, 1975] to show that strangers are concerned with increasing the predictability about the behavior of both themselves and others in the interaction that occurs when they first meet. Uncertainty reduction theory has been used in research on dating sites to explain how much information may be shared in profiles [Gibbs et al., 2010], using self-disclosure as the process of making the self known to others [Derlega et al., 1987]. According to this approach, people should disclose as much information in their host profiles as they feel comfortable sharing — a directive that presents various challenges, including the risk of over-exposure if the profile is public [Gibbs et al., 2010, Bazarova and Choi, 2014]. Nevertheless, the more information disclosed in a profile, the more likely it is perceived as trustworthy.

According to the Profile as Promise framework, one way to understand what kinds of information people disclose in the kind of static and asymmetric context of online profiles is signaling theory [Ellison and Hancock, 2013], which considers the relationship between explicitly stated signals and the underlying qualities they are likely to represent [Donath, 2007, Spence, 2002]. Spence's original application of signaling theory [Spence, 2002] explored how potential employees in a labor market tried to convey attributes that cannot be observed directly, such as reliability or goodness-of-fit with a company's culture. Signaling theory was also concerned with how such signals can be assessed, for example, by the employer. Some signals, called conventional signals, are relatively easy to fake; such as asserting that one is a reliable and hard worker, when in fact

one is not. Other signals, called assessment signals, are more difficult to fake; such as claiming that one has a degree from a prestigious institution. Gambetta used signaling theory to study how taxi drivers [Gambetta and Hamill, 2005] and criminals [Gambetta, 2009] form impressions and assess the trustworthiness of other parties based on cues, especially in fleeting initial interactions where high uncertainty is involved. Here we examine how Airbnb hosts signal their trustworthiness in their profile through self-disclosure and how these signals are perceived.

4.3 Step 1: Self-Presentation and Perceived Trustworthiness of Airbnb Host Profiles

Emerging literature is examining how people assess trustworthiness through self-disclosures made in online profiles. The *Profile as Promise* [Ellison and Hancock, 2013] conceptual framework, for example, incorporates the risks and rewards associated with assessing signals in a profile for whether said profile's promises can be trusted. Researchers had examined how individuals produce and assess trustworthiness signals in online dating profiles [Toma et al., 2008] and in online résumés [Guillory and Hancock, 2012]. However, we still know very little about what people self-disclose, and how that information is evaluated for trustworthiness in the context of sharing economy platforms such as Airbnb.

Given the importance of profiles, in this work we aim to advance our understanding of the type of content Airbnb hosts self-disclose in their profiles, and to determine the impact of these disclosures on perceived trustworthiness and host choice. We build on the Profile as Promise framework [Ellison and Han-

cock, 2013], drawing on theories from economics and communication to predict what kinds of information hosts will disclose in their profiles, and what kinds of disclosure will enhance trust. In particular, we apply uncertainty reduction theory [Berger and Calabrese, 1975] to predict that both quantity and diversity of information increases the perception of trust. We also draw on signaling theory [Donath, 2007, Spence, 2002] to predict that specific kinds of information can signal trustworthiness in a profile.

Specifically, we use a mixed-methods approach, with qualitative analysis, large-scale annotation, and an online experiment to examine the text-based self-descriptions of Airbnb host profiles. We qualitatively develop a categorization scheme that characterizes the primary self-disclosure topics in these profiles. We then quantitatively show how predictions from uncertainty reduction theory and signaling theory apply to this case, revealing that an increase in the quantity of content and the inclusion of specific topics can enhance perceptions of trustworthiness. Finally, we use an online experiment to show that the perceived trustworthiness of profiles is a significant predictor of host choice. Our results have practical design implications for platforms facilitating social exchange in the sharing economy.

Research Questions

In our examination of Airbnb profiles, our first objective, according to the Profile as Promise framework, is to determine what kinds of information hosts provide in their profiles to reduce uncertainty and signal trustworthiness.

RQ 1. What kinds of information do hosts self-disclose to signal their trustworthi-

ness?

The second question is concerned with how the information hosts disclose in their Airbnb profiles translates into perceived trustworthiness by guests. On Airbnb, like other peer-based sharing economy services, signals of trustworthiness are particularly important as trust is critical for social exchange [Cheshire and Cook, 2004, Lampinen and Cheshire, 2016]. Trust on Airbnb and other online marketplaces is tied to ratings and reputation. However, ratings on Airbnb tend to not be informative as they are skewed high, and there is initial evidence that the number of reviews received is predictive of room sales even when controlling for scores [Lee et al., 2015a, Zervas et al., 2015]. Research suggests that profile information matters on Airbnb: in one study, profile images were linked to the perceived trustworthiness of hosts and higher prices [Ert et al., 2016]. At the same time, the study showed that online review scores had no effect on the listing price, although profile text was not considered. We therefore ask:

***RQ 2.** What is the effect of different types of self-disclosure on perceived trustworthiness?*

Note that it is not immediately clear that trustworthiness maps directly to the choice of host. Choice can clearly be influenced by other factors, such as assurance [Cheshire, 2011]. There is initial data-driven evidence that visual-based trustworthiness impacts choice [Ert et al., 2016], even when reputation scores are manipulated to increase their variance.

Given the difficulty in establishing the causal link between profile disclosures and guest decision-making, we conduct an experiment that isolates the trustworthiness of a profile's disclosures from other external factors — such as reputation

indicators — and manipulates the effect of low versus high trustworthiness profiles on a decision-making task. In this experiment, we focus on addressing the following research question:

RQ 3. Do profile-based perceptions of trustworthiness predict choice of host on Airbnb?

4.3.1 Study 1 — How do Hosts Self-Disclose?

The primary goal of Study 1 is to uncover what Airbnb hosts self-disclose in the text-based self-descriptions in the profile. To the best of our knowledge, there is no established coding scheme for self-disclosure in this particular context. For this reason, Study 1 uses qualitative methods to develop, validate and apply a coding scheme for self-disclosure in Airbnb host profiles. We accomplished this task using a two-phase approach. In phase one, we developed and validated a coding scheme for topics of self-disclosure by qualitatively analyzing Airbnb profiles and using an inductive and iterative approach to identify categories. In phase two, we applied the coding scheme to a large set of host profiles, and examined patterns of self-disclosure on Airbnb.

Phase 1 — Developing and Validating Coding Scheme

Step 1: Development

To create the self-disclosure coding scheme for Airbnb, we used an iterative, inductive analysis for content-topic categories of information in host profiles. As this study is exploratory in nature, we established some guidelines for developing

the initial coding scheme. In particular, when creating an Airbnb profile page, the website prompts the host to share a few details about themselves, calling out three types of self-disclosure: “things you like”, “style of traveling or hosting”, and “life motto”. We were cognizant of the Airbnb interface prompt and used it as a starting point, fitting codes to the prompt topics and refining them according to the content, but not restricting our coding to topics suggested by these prompts.

For this step, we constructed a *Development Dataset* consisting of 300 sentences randomly drawn from a weighted sample of 203 host profiles from 12 major U.S. cities. The profiles were extracted from an open-sourced Airbnb dataset collated by an independent organization, Inside Airbnb [Inside Airbnb., 2016]. Non-English profiles were filtered out. We provide more details of the full dataset below.

Two raters independently coded the topics in each of the 300 sentences in the *Development Dataset*, using the qualitative data analysis and research software Atlas.ti. In addition to Airbnb prompts, the coders also considered topics used in previous self-disclosure studies [Jourard and Lasakow, 1958, Ma et al., 2016, Rubin, 1975]. After a full round of independent coding, the two coders compared their codes and deliberated, further merging the codes into concepts and topics. This analysis and coding process resulted in nine initial topic categories.

Step 2: Adjustment and Validation

In this step, we evaluated, adjusted, and validated the coding scheme for reliability and coverage, i.e. the percentage of sentences our codes applied to. In order to apply the coding scheme to a large set of host profiles on Airbnb, we designed a web interface to recruit annotators from Amazon Mechanical

| Topic | Agreement | | | Description |
|---------------------|-------------|-------------|-----------|---|
| | Vote- R1 | Vote- R2 | R1- R2 | |
| Interests & Tastes | .77 | .85 | .78 | Favorite books, music, hobbies, how I spend weekends and evenings, favorite ways of spending spare time. |
| Life Motto & Values | .51 | .56 | .52 | Life motto, values, philosophies; e.g. "Live courageously, love passionately". |
| Work or Education | .83 | .86 | .79 | Current or past job, school, major; e.g. "I'm an architect and designer". |
| Relationships | .69 | .62 | .60 | Family, significant other, pet; e.g. "I have a beautiful 16 year old daughter, a little sweet terrier Nora, two fish & a frog." |
| Personality | .80 | .70 | .65 | e.g. "I am extremely down to earth and I am a self-diagnosed work-a-holic". |
| Origin or Residence | .78 | .69 | .76 | Where from, current residence, history of moving; e.g. "I lived in D.C. for 5 years and Philly for 2 years"; "We both really love how much Chicago has to offer." |
| Travel | .78 | .72 | .83 | Love for travel; past travels; favorite travel destinations. |
| Hospitality | .73 | .54 | .66 | Welcoming or greeting, reasons for hosting, demonstrating availability; e.g. "We're delighted to be your hosts and tour advisors during your stay here." |

Table 4.1: Topics of self-disclosure in Airbnb host profiles.

Turk (AMT). The web interface presented each individual sentence from the host profiles, together with the initial topics and descriptions (some with examples) that were developed through the coding process in the foregoing step. The annotator was instructed to tag all topics that appeared in the sentence (a sentence could mention multiple topics). If none of the topics applied, the annotator was instructed to choose "other".

We validated the reliability of the coding scheme using two metrics: the level of agreement among crowd workers, and the level of agreement between the crowd workers' consensus and expert annotations, i.e. researchers from our

team.

To compute the first metric, the level of agreement among crowd workers, we constructed the *Initial Validation Dataset*, consisting of 300 sentences drawn from a new sample of 203 profiles. Sentences in the *Initial Validation Dataset* did not overlap with those in the *Development Dataset*. We recruited crowd workers from AMT to annotate sentences in the *Initial Validation Dataset* using a web interface that we developed (paying \$.02 per annotation), and computed Fleiss' kappa among the workers. There were four topics that had a Fleiss' kappa score lower than 0.5, indicating an unsatisfactory level of internal agreement. We iterated on the initially developed set of topics to address this issue, adjusting the name and description of two topics, and merged two closely related topics into one ("Hosting Attitude" and "Hosting Action" to "Hospitality"). After the edits to the coding scheme, eight topics remained, shown in the first column of Table 4.1.

To compute the second metric — the level of agreement between the consensus from crowd workers and expert annotations — we constructed the *Final Validation Dataset*, consisting of all 871 sentences from the text of a new batch of 203 profiles. The new profiles did not overlap with either those in the *Development Dataset* or those in the *Initial Validation Dataset*. Again, we asked three crowd workers to annotate each sentence, and used a majority voting rule to produce the final *vote* across three workers. A topic label for a sentence was retained only if at least two out of three workers indicated that the sentence mentioned that topic.

In terms of coverage, in total, at least two voters agreed on one or more topics for 91.5% of the 871 sentences (if we consider the workers' using an optional "other" category, a majority vote was achieved for 97.4% of the sentences). The

raters inspected the sentences where the workers did not reach an agreement, and verified that they did not contain significant missed themes.

Finally, two raters coded a 300-sentence sample from the *Final Validation Dataset*. We computed the agreement of these three different sources by calculating the pairwise Cohen's kappa scores for the worker majority vote (*vote* in Table 4.1), Rater 1 (*R1*), and Rater 2 (*R2*). The results of each pairwise agreement computation are shown in Table 4.1. The results suggest moderate to almost perfect agreement across all topics, and indicate that the coding scheme (the topic names, descriptions, and the set-up of majority votes from AMT) is reliable.

Phase 2 — Applying the Coding Scheme to Profiles

With the coding scheme validated, we could now annotate a large set of host profiles and examine the trends of self-disclosure. What might we expect for the disclosures? According to the Profile as Promise framework [Ellison and Hancock, 2013], hosts should disclose information they believe will signal to potential guests that they will be a trustworthy host. These disclosure goals should cause hosts to prioritize the disclosure of information that enhances trustworthiness. Signaling theory further suggests that perceptions of trustworthiness may be affected by the kind of signal [Donath, 2007, Spence, 2002]. If this is the case, then hosts should disclose more assessment signals (i.e. disclosures that can be verified):

H1.1 Hosts will disclose more about categories that have more assessment value, including Work or Education, and Origin or Residence, than about categories that have more conventional value, including Interests & Tastes, and Personality.

The Profile as Promise framework also suggests that information should be disclosed about the most relevant underlying qualities that the host is promising to potential guests. In this context, the type of hosting situation, on-site versus remote, should lead to different disclosure patterns. The on-site hosts (who share their space with guests during their stay) need to signal what kind of person a guest might meet. The remote hosts (who are not present) need to signal that the guests will be taken care of in their absence. Previous work on Airbnb revealed that on-site versus remote hosting is an important part of sociability within the host-guest relationship [Ikkala and Lampinen, 2015, Lampinen and Cheshire, 2016]. When hosting on-site, guests and hosts may have more substantial face-to-face interaction.

Given the increased likelihood of social interaction for on-site hosts, there is uncertainty about whether the guests and hosts will get along. We can draw on uncertainty reduction theory [Berger and Calabrese, 1975] to predict that on-site hosts will disclose more information than remote hosts in an effort to reduce the uncertainty for potential guests given that guests and hosts will socially interact. In particular, on-site hosts should disclose more information relevant to relationship development, such as one's preferences and personality.

H1.2 On-site hosts will disclose more, especially for topics that can reduce uncertainty during the interaction of sharing spaces, such as Interests & Tastes, and Personality, than remote hosts.

We first report on the dataset of Airbnb profiles we used for this analysis and throughout the rest of this work. Then, we describe the process of applying the coding scheme to annotate a larger portion of the host profiles. Finally, we discuss the results of testing the hypotheses using the annotated data.

Airbnb Dataset

To apply the coding scheme to a larger portion of Airbnb profiles, we used the large-scale dataset collected by Inside Airbnb [Inside Airbnb., 2016]. Inside Airbnb periodically scrapes the Airbnb website, making snapshots of Airbnb listings from 35 cities in 13 countries (at the time of writing) available for download. For each city, Inside Airbnb conducted a URL query through Airbnb search and scraped all public listings. We manually examined 10 samples from five cities in the dataset, visiting the Airbnb website for each entry to verify that the scraped data is consistent with the actual listing. Since each listing is always associated with a host, the free text portion of the host profile is available from the Inside Airbnb dataset. Some other metadata about hosts, in addition to the host self-description (the focus of the present work) include: host ID (a unique identifier for a host across the Airbnb platform), first name, type of listing (Entire Home/Apt, Private Room, or Shared Room), and whether the host is a “superhost” on Airbnb.

We limited our scope of analysis to U.S. and English-language host profiles only. Host profiles from other countries may contain non-English phrases or characters, introducing sources of noise, and making it difficult for crowd workers to annotate. We performed source language detection using the Google Translate API [Google, 2016] for each sentence in a host profile, and filtered out those containing non-English sentences.

The Inside Airbnb data included 93,361 listings across 15 U.S. cities. We first de-duplicated hosts from multiple listings by host ID. We used data from the 12 largest cities, excluding 3 cities with fewer than 1,000 unique hosts (Asheville, Oakland, and Santa Cruz County). We verified that the exclusion of these three

cities did not affect the results from Study 1. For the remaining 12 cities, we deduplicated 89,965 listings to obtain 67,465 unique hosts. Out of these unique hosts, we further filtered out 20,710 (or 30.7% of the de-duplicated quantity) host profiles with empty self-descriptions, and 6,750 (10.0% of the de-duplicated quantity) that contained non-English phrases.

In the end, we had 40,005 non-empty, English-only unique host profiles from 12 U.S. cities, with the following breakdown: New York (data collected in Sep 2015; 14,513), Los Angeles (Jan 2016; 8,062), San Francisco (Nov 2015; 3,400), Austin (Nov 2015; 2,477), Chicago (Oct 2015; 2,149), Seattle (Jan 2016; 1,798), Washington D.C. (Oct 2015; 1,633), San Diego (Jun 2015; 1,522), Portland (Sep 2015; 1,415), New Orleans (Sep 2015; 1,173), Boston (Oct 2015; 922), and Nashville (Oct 2015; 941).

With our previously validated coding scheme, we annotated the topics of a larger portion of host profiles from the above-mentioned Airbnb dataset using AMT, following the exact same procedure as described for annotating the *Final Validation Dataset*. We constructed the *Annotation Dataset*, consisting of all 4,377 sentences from 1,031 profiles, randomly selected using a weighted sample according to the number of unique non-empty host profiles in each city. As the coding scheme was the same as that used for the *Final Validation Dataset*, we merged the results from the *Annotation Dataset* and the *Final Validation Dataset*, forming the *Experiment Dataset* to boost the amount of annotated data, resulting in 5,248 annotated sentences from 1,234 profiles.³

Self-Disclosure Trends

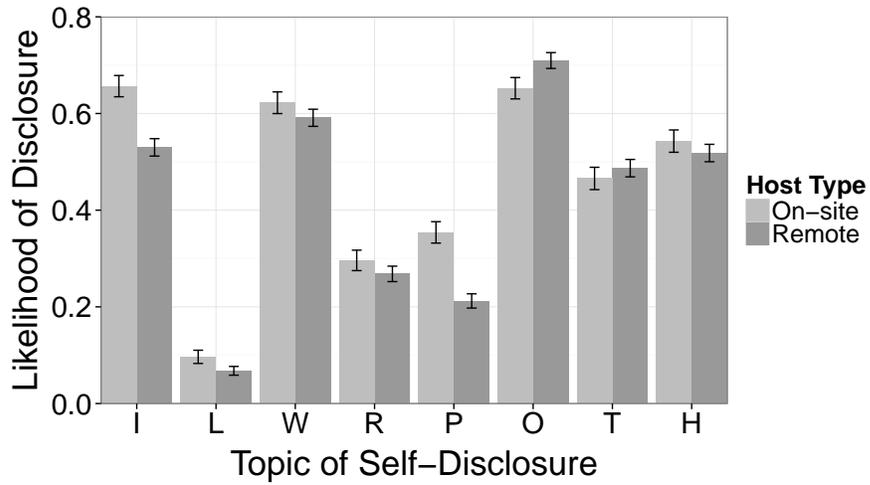
³The *Experiment Dataset* and other data used in this work are available from <https://github.com/sTechLab/AirbnbHosts>.

What do Airbnb hosts self-disclose in their profiles? We found that hosts were most likely to talk about *Origin or Residence* (68.8%), followed by *Work or Education* (60.29%) and *Interests & Tastes* (57.78%). There was substantial travel-related disclosure including writing about *Travel* (47.89%) and demonstrating *Hospitality* (52.76%). The topics that were less commonly mentioned were *Relationships* (27.88%), and *Personality* (26.58%). The topic that was least mentioned was *Life Motto & Values* (7.86%).

This pattern of results is partially supportive of H1.1 and the prediction from signaling theory that hosts would disclose more assessment signals than conventional ones. Consistent with the hypothesis was the frequent disclosure of assessment signals regarding *Origin or Residence* and *Work or Education*, and the low levels of disclosure regarding *Personality*. The frequent disclosure of *Interests & Tastes*, however, did not line up with the hypothesis. The analysis below on disclosures by host type provides some insight: the high rate of disclosure of *Interests & Tastes* was driven in part by on-site hosts, which may have been part of an effort to reduce uncertainty for guests who would be meeting their hosts.

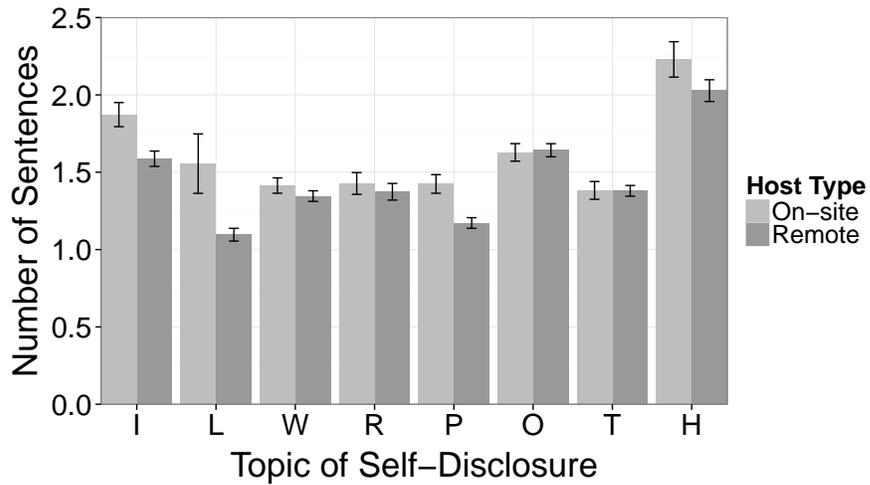
Self-Disclosure Trends by Host Type

Addressing H1.2, we compared the self-disclosure of hosts based on the type of property they offered: on-site versus remote. As hypothesized, on-site hosts ($M=66.12$, $SD=59.66$) on average wrote longer profiles (measured by word count) compared to remote hosts ($M=55.85$, $SD=52.09$), $t(880)=3.07$, $p < .01$. Second, in terms of topics, we found that on-site hosts were more likely than remote hosts to write about topics that signal their personality and tastes. We calculated the percentage of profiles that mentioned each topic for different host types, shown in Figure 4.1. For each topic of disclosure (identified by the first



.5

Figure 4.1: Probability of disclosure per topic



.5

Figure 4.2: Average number of sentences per topic

Figure 4.3: Self-disclosure trends by topic and host type. The error bars represent one standard error.

letter on the x -axis), we show the percentage of profiles of each host type that mentioned that topic (y -axis). As the figure reveals, on-site hosts were more likely to write about topics of *Interests & Tastes* ($\chi^2=18.57$, $df=1$, $p < .001$) and *Personality* ($\chi^2=29.18$, $df=1$, $p < .001$); and less likely to mention *Origin or Residence* ($\chi^2=4.17$, $df=1$, $p < .05$). Note that the results for *Interests & Tastes* and *Personality* remain statistically significant with Bonferroni correction for multiple tests.

We also compared the *number of sentences* used for the different disclosure topics, which shows similar trends (Figure 4.2). For *Interests & Tastes*, on-site hosts on average wrote more sentences ($M=1.87$, $SD=.08$) than remote hosts ($M=1.59$, $SD=.05$), $t(535)=3.09$, $p < .01$. For *Personality*, we also see that on-site hosts on average wrote more sentences ($M=1.42$, $SD=.06$) than remote hosts ($M=1.17$, $SD=.03$), $t(260)=3.64$, $p < .001$. Finally, on-site hosts on average wrote more sentences mentioning *Life Motto & Values* ($M=1.56$, $SD=.19$), $t(48)=2.34$, $p < .05$ than remote hosts ($M=1.10$, $SD=.04$). Again, the results for *Interests & Tastes* and *Personality* remain statistically significant with Bonferroni correction for multiple tests.

To rule out the possibility that these differences are due to a host's level of experience (e.g. hosts may modify their profiles to write about specific topics more as they host more guests), we conducted a similar analysis comparing average hosts with superhosts, a qualification type assigned by Airbnb [Airbnb, 2019] for hosts that meet several criteria, including frequent hosting, high response rate, and high review scores. We omit the details of this analysis for brevity, but note that despite the fact that superhosts wrote significantly longer profiles (a mean of 72.13 words compared to 57.74 words for non-superhosts, $t(220)=3.01$, $p < .001$), there was no significant difference between the groups

in the likelihood of mentioning *Interests & Tastes* or *Personality*. This analysis suggests that the difference between on-site and remote hosts is not due to any difference in experience of hosting, but rather due to the expected differences in the type of interaction that is going to take place. Taken together, the results support uncertainty reduction theory's central contention that people seek to reduce uncertainty in the face of new relationships.

4.3.2 Study 2 — Self-Disclosure and Perceived Trustworthiness

Study 1 revealed that we can classify host disclosures into eight topics, and that these topics can be reliably assessed by independent coders. An important next question is whether the hosts were disclosing information that enhanced perceptions of their trustworthiness. That is, do the topics that hosts disclosed the most in Study 1 lead to higher levels of perceived trustworthiness? To examine this question, we asked online participants to rate how trustworthy they found each profile.

One way to operationalize the concept of trustworthiness is by using three key dimensions: ability, benevolence and integrity [Mayer et al., 1995]. These three dimensions are closely related but may have different effects on trust depending on context [Colquitt et al., 2011]. Further, these dimensions are all likely to be relevant for Airbnb profiles. In the context of Airbnb, ability refers to domain-specific skills or competencies that the host has. Benevolence refers to the extent to which the host is believed to want to do good to the guest beyond profit-driven motives. Finally, integrity refers to the host adhering to a set of moral principles and rules.

How might the disclosures in Airbnb profiles influence these dimensions of trustworthiness? The Profile as Promise conceptualization of the profile as a psychological contract implies that information provided in the profile is an obligation by the host to a guest, namely that the information disclosed in the profile is trustworthy and will not misrepresent the host or the host's home [Ellison and Hancock, 2013]. This notion of the psychological contract suggests that hosts should be sensitive to how their promises will be evaluated for trustworthiness by potential guests. If this is the case, then hosts should produce promises that signal trustworthiness. We can draw on the same theoretical perspectives we used to characterize the production of disclosures to specify predictions about how the profile disclosures on Airbnb affect evaluations of trustworthiness. First, uncertainty reduction theory [Berger and Calabrese, 1975] predicts that the more information hosts disclose, the more the profile will reduce uncertainty for profile viewers, which should enhance how trustworthy they will be perceived. Note that more diverse information should lead to more uncertainty reduction. That is, profiles that disclose more kinds of information will be perceived as more trustworthy than profiles that simply say a lot about fewer things. We therefore predict that:

H2.1 Longer and more diverse self-disclosures are perceived as more trustworthy.

Secondly, Study 1 demonstrates that hosts communicate a variety of topics in their profiles. Signaling theory predicts that hosts use these disclosures to signal underlying qualities or attributes that should enhance the perceptions of their trustworthiness as a host. If the hosts have optimized their signaling behavior for trustworthiness, then the categories they disclose most often should be the categories of disclosure that are perceived as most trustworthy. Thus, profiles

| | |
|-----|---|
| A1. | This person is capable of paying his/her own rent or mortgage. |
| A2. | This person maintains a clean, safe, and comfortable household. |
| B1. | This person will be concerned about satisfying my needs during the stay. |
| B2. | This person will go out of his/her way to help me in case of an emergency during my stay. |
| I1. | This person will stick to his/her word, and be there when I arrive instead of standing me up. |
| I2. | This person will not intentionally harm, overcharge, or scam me. |

Table 4.2: Six-item perceived trustworthiness scale.

with disclosures that were observed frequently in Study 1, including *Origin or Residence, Work or Education, Interests & Tastes, and Hospitality* should be perceived as more trustworthy than profiles that do not contain these topics.

H2.2 Self-disclosure topics used most frequently by hosts will be associated with increased perceived trustworthiness compared to less frequent topics.

Methods

As mentioned above, we are interested in the perceived trustworthiness of host profiles. To measure trustworthiness, we developed a six-item perceived trustworthiness scale on three dimensions: ability, benevolence, and integrity [Mayer et al., 1995]. Based on items in the scale developed by Mayer et al. for an organization context [Mayer and Davis, 1999], we developed new items that measure trustworthiness in the context of hosting. These items are shown in Table 4.2. Items A1–A2 measure ability; items B1–B2 measure benevolence; and items I1–I2 measure integrity. When asking for profile ratings, these items were shown in a random order.

Procedure

To assess the perceived trustworthiness of host profiles, we recruited crowd workers from AMT to rate host profiles in the *Experiment Dataset* using the perceived trustworthiness scale. We split the *Experiment Dataset* profiles into batches of 20, and had five different workers annotate each batch. Recall that these profiles were already labeled with the topics. We used 1,200 of the 1,234 profiles in the *Experiment Dataset* for this study. For each profile, workers were instructed to rate their level of confidence regarding each of the statements, on a scale from 0 to 100, with steps of 10. The task required that each worker only rate one batch of the profiles to prevent any single worker's perception from being over-represented in the results. Workers were paid \$1.00 for each task.

At the beginning of each task, we used a paraphrase question borrowed from [Munro et al., 2010, Danescu-Niculescu-Mizil et al., 2013a] to check the linguistic attentiveness of each worker. We re-issued the task if we received an incorrect response to this question. To create the perceived trustworthiness score, we calculated the perceived trustworthiness as the mean of responses for all six items by the five workers that rated the same profile. For some analyses, we also used three trustworthiness dimensions separately, with each score calculated as the average of the two relevant items.

Results

We investigated the effects of profile length and diversity (H2.1), and topic (H2.2) on perceived trustworthiness. Generally, the mean ability score of the 1,200 profiles was 68.82, $SD=13.84$; the mean benevolence score was 63.94, $SD=13.97$; the mean integrity score was 66.79, $SD=13.37$. Note that perceived trustworthiness scores across the three dimensions were highly correlated [pairwise Pearson's R

correlation: A–B (initials): 0.86; A–I: 0.88; B–I: 0.92; $p < .001$].

Length, Diversity and Perceived Trustworthiness

To examine the effect of length (word count) on perceived trustworthiness, we plot the relationship between length (x -axis, log scale) and perceived trustworthiness (y -axis) on each of the three trust dimensions in Figure 4.4.

Supporting H2.1, Figure 4.4 shows a clear relationship between increased profile length and perceived trustworthiness scores. This relationship is confirmed by linear regression with log transformation for profile length [$b = 7.89$, adjusted $R^2 = .38$, $F(1, 1198) = 721.4$, $p < .001$]. This means that when a profile doubles in length, the perceived trustworthiness score increases by approximately 5.47, suggesting a pattern of diminishing returns when hosts write longer profiles.

To illustrate this pattern, we divide the profiles into deciles and calculate the average perceived trustworthiness score for each decile. Comparing profiles in the second decile (mean word count: 13) to those in the first (mean word count: 6), mean trustworthiness score increased 18.9%; whereas comparing profiles in the ninth decile (mean word count: 106) to those in the tenth (mean word count: 188), mean trustworthiness score only increased by 2.5%.

H2.1 also predicts that, in addition to overall length, the number of topics will also have a positive impact on trustworthiness scores. We performed multiple linear regression analysis with the number of topics, length as control, and the interaction length \times number of topics [adjusted $R^2 = 0.39$, $F(3, 1196) = 256.6$, $p < .001$]. The analysis showed that the number of topics contributes to perceived trustworthiness [$b = 4.47$, $t(1198) = 5.54$, $p < .001$] even when controlling for length [log scale, $b = 9.53$, $t(1198) = 15.03$, $p < .001$]. There was also an interaction

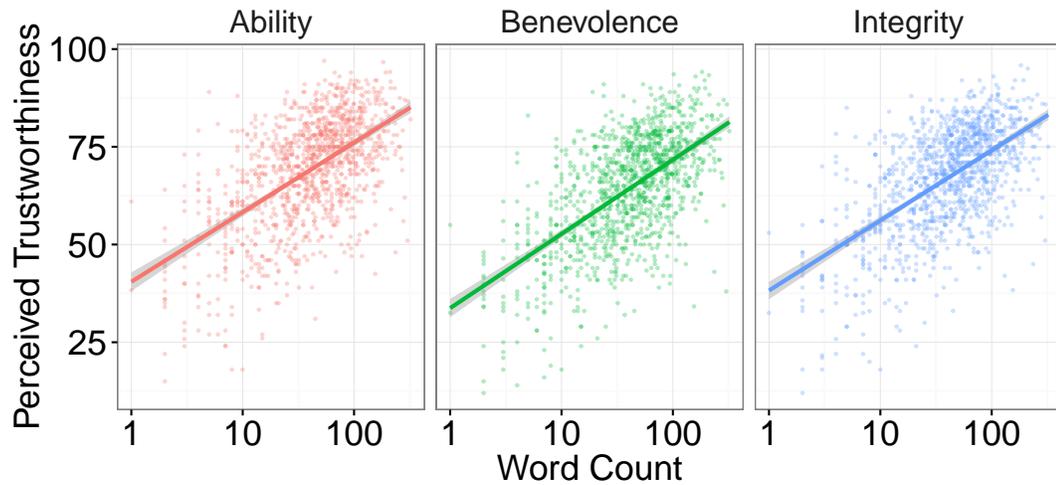


Figure 4.4: Perceived trustworthiness increases with profile length (x -axis on log scale).

effect between length and topic count [$b = -0.95$, $t(1198) = -5.04$, $p < .001$], indicating that for shorter profiles, the number of topics increased perceived trustworthiness even more.

Figure 4.5 visualizes the relationship between perceived trustworthiness and the number of topics mentioned in the profile. Each line represents the density distribution of perceived trustworthiness scores for profiles that mention a fixed number of topics. For example, the darkest lines represent the distribution of perceived ability, benevolence, and integrity of one-topic profiles. The figure shows that there is variation in trustworthiness score within each topic count bin, but as topic count increases, the trustworthiness scores also increase, and the variations become smaller. Note that here we are *not* showing the effect of profile length, which was illustrated in Figure 4.4.

Topic and Perceived Trustworthiness

We now analyze the effect of topic choice on trustworthiness scores. Recall that H2.2 predicted that the topics disclosed most frequently by hosts in Study 1, namely, *Origin or Residence*, *Work or Education*, *Interests & Tastes*, and *Hospitality*, would also be evaluated as most trustworthy.

In our dataset, there were eight profiles that did not mention any topics, 117 one-topic profiles, 231 two-topic profiles, 239 three-topic profiles, 269 four-topic profiles, and 336 profiles that mentioned five or more topics. We focus on profiles that are limited to one-topic, two-topic, and three-topic combinations. For example, looking at two-topic combinations, there are $8 \times 7/2 = 28$ different options, although, as we show below, there are some topic combinations that are more common than others. These 1-3 topic combinations have the most variation, but are also simpler to study, as understanding the impact of one single topic amid all combinations of different sizes is highly unlikely even with 1,200 profiles. We call these different combinations of topics “strategies”, and compare the relative success of different strategies controlling for the number of topics.

We have shown that as the number of topics increase, the trustworthiness scores also increase. We computed one-way ANOVAs comparing the *relative* effectiveness of strategies within each of the one-topic, two-topic, and three-topic profile groups. For one-topic profiles, there was a significant effect of strategy on ability [$F(7, 109) = 7.79, p < .001$], benevolence [$F(7, 109) = 8.55, p < .001$], as well as integrity [$F(7, 109) = 7.36, p < .001$]. For two-topic profiles, there was no significant effect of strategy on ability [$F(5, 225) = 1.54, p = .18$], but a significant effect on benevolence [$F(5, 225) = 3.95, p < .01$], as well as integrity [$F(5, 225) = 2.74, p < .05$]. For three-topic profiles, there was a significant effect

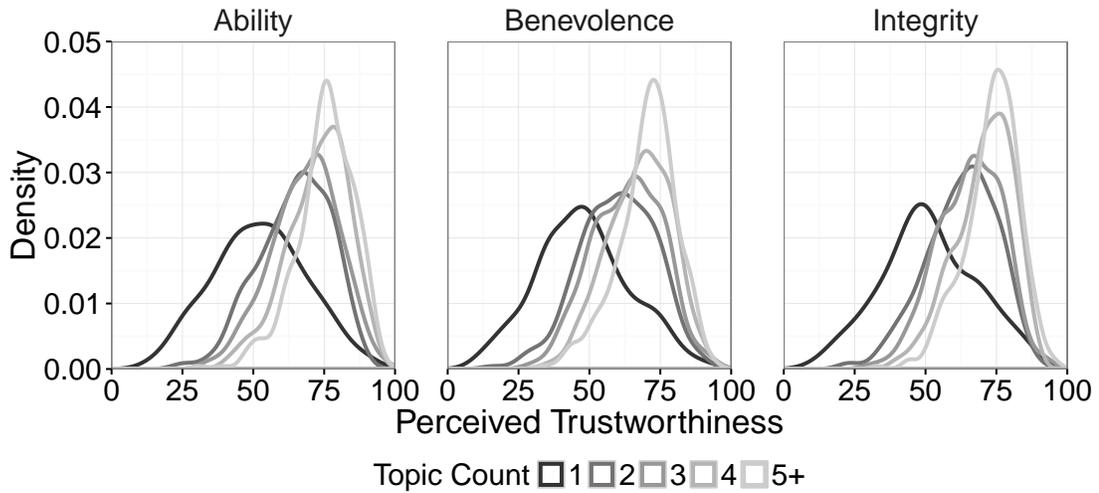


Figure 4.5: Perceived trustworthiness score distributions for profiles with different number of topics.

of strategy on ability [$F(8, 230) = 2.83, p < .01$], benevolence [$F(8, 230) = 4.21, p < .001$], as well as integrity [$F(8, 230) = 2.95, p < .01$].

Figure 4.6 shows the raw data for this analysis, organized by the number of topics (panels), and three dimensions of trustworthiness scores (columns). Every row is marked on the left with the initials of the topics in the self-disclosure strategies (e.g. in the first row of the second panel, *OW* stands for the topic combination of Origin or Residence, and Work or Education). On the right, we show the number of profiles using this strategy (54 for *OW*, the most of all two-topic strategies). The vertical lines in each row represent profiles, positioned at the value of the perceived trustworthiness score on each dimension. The color indicates whether the profile falls within the bottom (red), or top (green) quartile of the profile group that used the same amount of topics (the dotted lines indicate the bottom and top quartile boundaries).

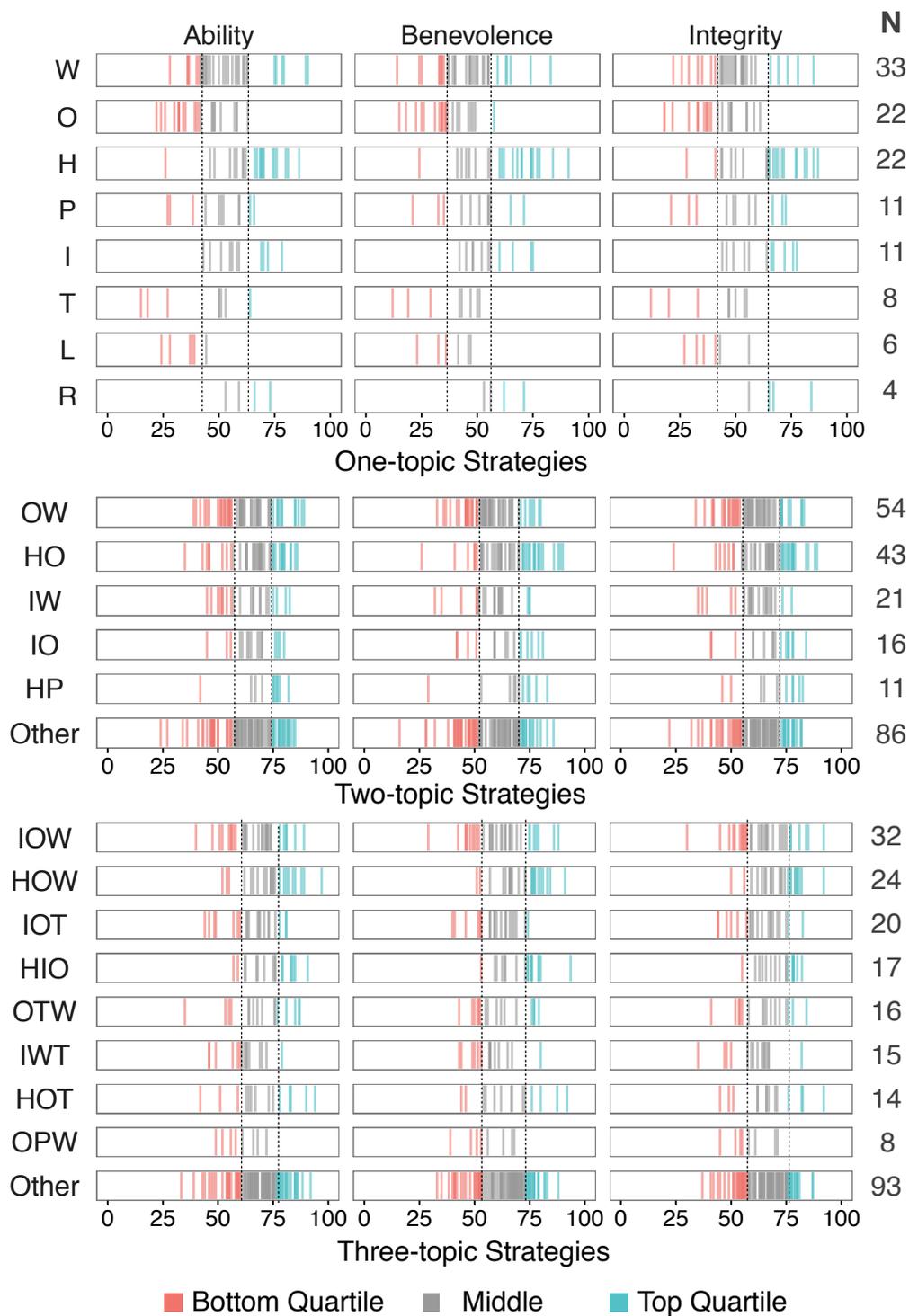


Figure 4.6: Comparison for different strategies, organized by the number of topics mentioned in the profile. The dotted lines indicate the bottom and top quartiles in each topic-count group.

There are several takeaways from Figure 4.6. Consider first the one-topic strategies: clearly, *Work or Education*, *Origin or Residence*) and *Hospitality* are the most popular, representing 66% of the one-topic profiles. Visually, it is clear from Figure 4.6 that the most successful single-topic strategy is *H*, where the profiles trending more to the right and top-quartile profiles appearing more frequently. This visual examination is confirmed with post-hoc comparisons using Tukey HSD tests. Among one-topic profiles, *Hospitality* was the best-performing strategy, significantly trumping *L*, *O*, *P*, *T* for ability; *L*, *O*, *T*, *W* for benevolence and integrity ($p < .05$, same for all the post-hoc comparisons reported henceforth).

The second-best one-topic strategy was *Interests & Tastes*, outperforming *L* and *O* for ability, and *O* and *T* for benevolence and integrity. Hosts were not very successful writing about *Life Motto & Values*, although Airbnb explicitly prompted them to, which underperforms *H*, *I*, *R*, *W* for ability; *H* for benevolence, and *H* and *R* for integrity. Finally, as reflected in Figure 4.6, *W* outperforms *O* for ability; *O* outperforms *R* for benevolence, *R* outperforms *O* and *T* for integrity. None of the other pairwise comparisons were significant.

Moving on to the two-topic strategies, the dominant strategies are *OW* and *HO*, both combinations of the most popular single-topic strategies *W*, *O* and *H*, covering 42% of two-topic profiles. Interestingly, the *WH* strategy was not often used (for three-topic combinations, *HOW* is again popular). The next two popular strategies are *IW* and *IO*, indicating that hosts add on *Interests & Tastes* as additional information. In terms of success for the two-topic strategies, post-hoc comparisons did not indicate significant differences among strategies for ability or integrity. However, for benevolence, *HO* outperforms *OW* and *Other*

(all other two-topic combinations that are not explicitly listed in Figure 4.6).

Finally, we see in three-topic combinations that the most common strategies are *IOW* and *HOW*.

In terms of relative success, post-hoc comparison indicated that *HOW* is clearly most successful, outperforming *IWT* for ability, *IOT*, *IOW*, *IWT*, *OPW*, *Other* for benevolence, and *IWT* for integrity. In addition, *HIO* outperforms *IOT* and *IWT* for benevolence. This may again be due to the high effectiveness of *Hospitality* as part of the disclosure strategy, when the host is making a direct promise to take care of the guests.

Overall, the pattern of results supports H2.2 and the prediction that profiles with topics most frequently disclosed by hosts are also those that are evaluated as most trustworthy. While hosts employed different combinations of topics as part of their self-disclosure strategies, it is clear that strategies that include the most frequently disclosed topics from Study 1 were the most successful in generating perceived trustworthiness: *Work or Education*, *Origin or Residence*, *Hospitality*, and *Interests & Tastes*. We now proceed to show that these trustworthiness scores are meaningful because they have a direct impact on host choice by potential guests.

4.3.3 Study 3 — From Perception to Choice

In this section, we examine how perception of trustworthiness leads to differences in host choice. As mentioned earlier, a number of factors may influence a potential guest's *decision* to stay with a host, such as availability, price, and characteristics of the property (e.g. location). Our primary question is whether

the trustworthiness signaled by profile disclosures can influence a potential guest's decision-making outcome, all other things being equal. To address this question, we isolate disclosures in the profile by conducting an online experiment to examine the effect of perceived trustworthiness on host choice. In particular, we vary the level of perceived trustworthiness, and test the extent to which the perceived trustworthiness of profiles influences a potential guest's choice.

Understanding choice has important real world implications. In the face of a potential social exchange opportunity with multiple exchange candidates, those who portray themselves as untrustworthy can potentially be "punished". As shown above, the content of an Airbnb host profile affects perceived trustworthiness. We know that trustworthiness differences can affect choice [Ert et al., 2016] in other settings, and hypothesize that:

H3.1 Higher perceived trustworthiness scores for text-based host profiles predict the likelihood of guest choice.

Methods

To test whether perceived trustworthiness affects a potential guest's decision-making, we employed a pairwise experiment to elicit guest response. Since we obtained a perceived trustworthiness score for each profile in Study 2, we paired profiles with different scores to examine if the score predicts guest's preference between two hosts in a pair. If the value of the trustworthiness score perfectly predicts choice, the observed pairwise decisions we obtain from respondents should follow the Bradley-Terry model [Bradley and Terry, 1952], which predicts the outcome of a comparison given associated values with each participant in

the match.

To this end, we generated profile pairs that were comparable in length, but with one high and one low perceived trustworthiness score. We controlled for length for a number of reasons. Firstly, we showed above that the length is highly correlated with trustworthiness. Choosing high- and low-scoring profiles from a global sample is therefore likely to result in unbalanced short and long profile pairs. We therefore used an adaptive matching method that takes length, then score into account.

First, we ranked 1,200 annotated host profiles based on word count. Then, from shortest to longest, we used a sliding window of roughly 240 profiles, with steps of size 120. All profiles within each window form a group. For each group, we calculated the bottom and top quartiles of mean trustworthiness score (the mean of the ability, benevolence, and integrity). We then iterated through every combination of two profiles, one from the bottom and one from the top quartile in that group, filtering out profile pairs where one profile is longer than the other by more than 20%. As the result of this process, we had 19,892 top-quartile-low-quartile (in the sliding window) profile pairs, representing 589 unique profiles.

The preference task for each profile pair was simple. First, each pair of profile descriptions was shown to a respondent. For the first five seconds, the profiles are shown but buttons were deactivated to encourage the respondent to read the profiles before making a decision. When the buttons become activated, the respondent click on one of the two profiles in response to the question,

“Which of the two hosts do you feel more comfortable staying with?” We

deliberately chose to ask about comfort, and not generally about host preference, e.g. “Which of the hosts would you choose to stay with?” Pilot studies we ran showed that, in making host preference decisions, people considered other personal and dyadic match factors, like their interest in staying in a location implied in the profile. While such considerations are generally interesting, the “comfort” phrasing was used to focus on the effect of the trustworthiness construct. A future study can address ecological validity by including other factors; we show that the trustworthiness construct *does* impact pairwise preference based on comfort, holding all else constant.

We recruited respondents from AMT for this task. Each respondent was asked to evaluate 50 profile pairs in each task. To improve the quality of the results, three pairs out of each batch of 50 were repeated in a random order, within each batch. If the respondent provided an inconsistent answer for more than one pair, we filtered their responses out of the analysis. Workers were paid \$1.00 per task.

Experimental Results

We obtained choice results from 423 unique respondents. Consistency filtering, and removing four responses with missing values, left us with 355 responses consisting of 16,685 pairs of choices, representing the 589 unique profiles.

We used the Bradley-Terry model [Bradley and Terry, 1952] to evaluate pairwise choice. The Bradley-Terry probability model predicts the outcome of a comparison given associated values with each participant in the match. If the perceived trustworthiness score accurately predicts choice, it should be a good fit to the theoretical Bradley-Terry model likelihood. Specifically, according to

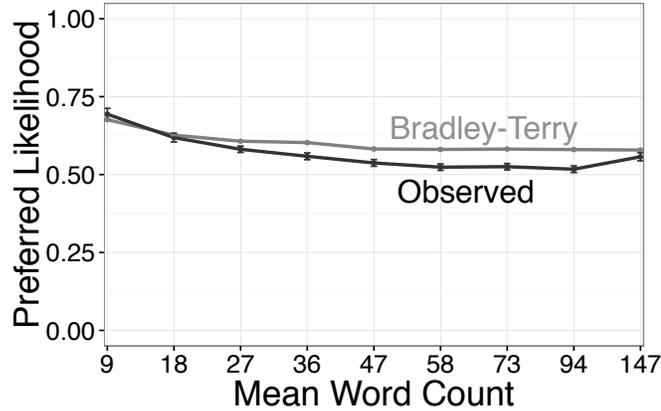


Figure 4.7: Observed likelihood that the host with high trustworthiness score is preferred and the probability predicted by Bradley-Terry model, by profile length. The error bars represent one standard error.

the Bradley-Terry model, the probability of profile i being picked compared to profile j is [Bradley and Terry, 1952]:

$$P(i \text{ is preferred to } j) = \frac{\lambda_i}{\lambda_i + \lambda_j} \quad (4.1)$$

where λ is a positive-valued parameter associated with each individual option. In our case, λ is the perceived trustworthiness score. Intuitively, the larger the difference between λ_i and λ_j , the higher the probability of i being chosen (with an upper-bound of 1). In addition, if the difference between λ_i and λ_j is fixed, the probability of i being chosen decreases as the absolute values of λ_i and λ_j increase (with a lower-bound of 0.5).

Figure 4.7 summarizes, for different buckets of profile length, the effectiveness of perceived trustworthiness in predicting choice compared to the prediction of the Bradley-Terry model. The x -axis shows the mean word count of the profile pairs in that length bucket, and the y -axis shows the likelihood of the profile with the higher trustworthiness score being chosen. Both theoretical (grey) and

observed (black) likelihoods are plotted. The figure shows that for profiles in shorter length groups (the first two groups on the left), perceived trustworthiness predictions closely match the Bradley-Terry probability. For longer profiles, however, the observed likelihood is lower than what is theoretically expected, indicating less predictive power. Nevertheless, even for the longer profile, as Figure 4.7 shows, the likelihood of choosing the top-quartile profile was higher than chance (50%), as shown by Exact-Binomial tests for each length group ($p < .05$) except for the eighth (the bin with mean word count of 94, $p = .06$).

The divergence of the longer profiles from the model may be due to at least two factors. First, and more mundanely, we ran the task on AMT where workers may not be incentivized to spend time reading longer profile descriptions, skewing the results towards random chance. Second, the results may reflect the fact that higher trustworthiness scores for longer profiles are not as predictive for decisions. With longer profiles, other factors are more likely to be mentioned, such as interests and tastes, that may generate specific dyadic attractions, and play a more significant role in influencing choice. We discuss this possibility in Section 4.5.

4.4 Step 2: A Computational Approach to Perceived Trustworthiness of Airbnb Host Profiles

Again, this chapter is focused on both gaining a deeper understanding of how people use language cues to establish trust in sharing economy, as well as predicting trust automatically. Step 1 has uncovered patterns on how host present themselves on Airbnb (eight common topics; longer profiles for on-site hosts

compared to remote hosts). In addition, initial understanding has also been established on what strategies people use to self-disclose and their relative effectiveness. However, due to the limitation in data size in Step 1, we did not attempt to build computational models to *predict* perceived trustworthiness directly. In Step 2, I scale up the data collection and develop computational models to predict perceived trustworthiness.

Specifically, we first enlarged the previous dataset to include 4,180 host profiles annotated with perceived trustworthiness scores, the largest such dataset to date. To enable the computational analysis, we developed several models building on various language-based features. Using these features and models, we evaluate a prediction task distinguishing profiles of low and high perceived trustworthiness. In addition, we use Lasso regression to examine the factors that contribute to higher and lower perceived trustworthiness. We discuss the results in relation to previous research on deception [Toma and Hancock, 2012] and loan defaults [Netzer et al., 2016], showing that the linguistic features contributing to higher perceived trustworthiness may not always align with features that were reported to be associated with other factors that may be indicative of *actual* trustworthiness.

This part of the project builds on a number of studies using computational approaches to study language and social interactions, enabled by new corpus and techniques from natural language processing and machine learning. For example, [Danescu-Niculescu-Mizil et al., 2013b] detects politeness computationally from text by constructing new datasets of online requests; [Mitra and Gilbert, 2014] found that in crowdfunding sites, language that makes direct promises such as “project will be” and “pledgers will receive” is predictive of a project being

funded; and finally, [Netzer et al., 2016] found that language in loan requests can help predict loan defaults.

4.4.1 Method and Dataset

Our main dependent variable here is the *perceived trustworthiness* of host profiles in the context of online lodging marketplaces. Trustworthiness is defined as an attribute of a trustee (the host). We measure *perceived* trustworthiness of hosts based on their profile texts alone through a custom scale developed in [Ma et al., 2017a], which asks potential guests about how confident they are that the host in question is capable, benevolent, and with integrity [Mayer et al., 1995]

We used the same source dataset as well as the annotating procedure described in Step 1. In other words, we used the Airbnb dataset collected by an independent organization, Inside Airbnb⁴ on 12 U.S. major cities, and conducted a weighted sample of profiles across all cities for 3,000 unique host profiles. We divided profiles into 150 batches, each containing 20 profiles, and recruited three annotators for each batch from Amazon Mechanical Turk (AMT), paying \$1.5 for each task. We required annotators to be based in U.S., adult, and with previous approval rate of at least 90%. We also required that each worker to only perform the task once (i.e. rate exactly 20 profiles). In this part of the project, we reduced the number of annotators per-profile from five in Step 1 to three, based on analysis showing that three raters can reach a satisfying level of inter-annotator agreement compared to five. Finally, in the exit survey, we collected additional information about the annotators that was not collected by [Ma et al., 2017a], including demographic information as well as their generalized trust attitude

⁴<http://insideairbnb.com/>

using the scale by [Yamagishi and Yamagishi, 1994].

We received 450 responses from the AMT workers and applied a series of filtering process to exclude potential “spammers”, including checking the answer to a linguistic attentiveness question, the standard deviation and mean of the ratings of the same worker, and task completion time. We filtered out responses that have very low and high (bottom and top 2.5%) standard deviation, mean, and task completion time. We retained the rating of a host profile if it has at least one rating after the filtering. In the end, we retained new annotations of perceived trustworthiness of 2,980 host profiles.

After initial filtering, we performed z-score standardization on the scores given by the same annotator, as we expected that each annotator’s scores are subjective with different baselines for trust. Indeed, our data shows a significant correlation between an annotator’s reported generalized trust attitude, and the average trustworthiness scores the annotator assigned to the 20 profiles [$\beta = .34$, $t(380) = 7.08$, $p < .001$], further justifying the decision to standardize the scores per annotator. After standardization, we took the average of scores given to the same profile by workers to be the perceived trustworthiness score of the profile.

To evaluate the reliability of annotations, we calculated the mean pairwise Pearson correlation for all profiles among three raters, pooling all the data. The average pairwise correlation is 0.49 (0.29 before standardization). Naturally, trustworthiness is a subjective concept. Our data showed patterns similar to data from previous research on politeness [Danescu-Niculescu-Mizil et al., 2013b]: higher agreement at the extremes and lower agreement in the middle, which motivated our evaluation setup as detailed below.

Since our annotation process is almost exactly the same as in Step 1, we merged the new dataset with the one reported in previous work in order to boost the amount of data available for training and testing. We performed z-score standardization on the previous dataset before merging with the dataset we newly acquired. As a result of this merging process, we now have an extended Airbnb host profile dataset containing a total of 4,180 profiles. The perceived trustworthiness scores in our extended dataset have a mean of zero and standard deviation of 0.8. We use this extended dataset in subsequent analysis. The extended dataset is available online⁵ and contains all profile texts, perceived trustworthiness annotation, as well as the demographic information and generalized trust attitude of annotators.

With the extended dataset, we set up two tasks: prediction and regression. For the first task, our goal is to find the best model that predicts perceived trustworthiness. As trustworthiness is a subjective concept, we set up the task as a binary classification, following the example in [Danescu-Niculescu-Mizil et al., 2013b]. We used logistic-regression classifiers and only top and bottom quartile of the profiles in different profile lengths buckets in terms of perceived trustworthiness score. For the second task, our focus is on understanding, which we address using Lasso regression for feature selection, also using the top and bottom quartile of the data in each length batch.

4.4.2 Predicting Perceived Trustworthiness

In this section, we set up the prediction task, and discuss features that we construct from profile text as inputs for different prediction models, as well as

⁵<https://github.com/sTechLab/AirbnbHosts-Extended>

model performance.

Evaluation Setup

We split our data into two parts: a training and cross-validation set (80% of data) that we use during model tuning, and a held-out set (the rest 20%) that is kept separate and reserved for the final test.

We frame the prediction task for profiles of different lengths. We know that length plays a significant role in predicting perceived trustworthiness [Ma et al., 2017a], which is again confirmed with our extended dataset [$\beta = .65$, $t(4, 178) = 55.61$, $p < .001$]. To this end, after trimming the outliers, i.e. the shortest and longest profiles (bottom 5% and top 5% in terms of word count), we divided the rest of the profiles into five equal batches based on word count. The batches, from shortest to longest profiles, have the following ranges of word count: 6–19 words, 20–36, 37–58, 59–88, and 89–179.

Within each batch, we calculate the bottom and top quartile of the perceived trustworthiness score. We then use logistic-regression classifiers to predict, for profiles in these two quartiles, whether they will be in the bottom quartile (zero), or top quartile (one), therefore only using 50% of the data in the cross-validation set. We measure the quality of our prediction using the accuracy of the classifiers (we are not using F1 and AUC scores as the labels are balanced).

Model Features

For our prediction models, we used different combinations of the features described below. We first performed the following data pre-processing. After removing punctuation and numbers using regular expression matching, we converted the remaining letter words into lowercases, and removed stop words using a union of lists of English stop words from NLTK and *scikit-learn* feature extraction module, consisting of 352 stop words. Finally, we lemmatized verbs and nouns using NLTK WordNet lemmatizer.

LIWC Features. We extracted 73 features from raw profile text (before pre-processing) using LIWC (Linguistic Inquiry and Word Count). LIWC is a dictionary-based text analysis tool that counts the percentage of words that reflect linguistic process, psychological process, and personal concerns. LIWC has been shown to predict numerous psychological outcomes [Pennebaker et al., 2001]. We used the 2007 version of LIWC and substituted readability in LIWC with Flesch-Kincaid grade level (extracted using the Python package *textstat*).

Bag-of-Words

We vectorized each of the pre-processed profiles using *CountVectorizer* from the *scikit-learn* library. We used one-, bi-, and tri-grams and required the grams to have appeared at least 20 times. This process resulted in 1,012 word features, which we use in our baseline model.

Sentence Categories

In Step 1, we manually developed a set of eight sentence categories (shown in the first column in Table 4.3) that frequently appear in Airbnb host profiles.

| Category | Accuracy | F1-Score | AUC |
|---------------------|----------|----------|-----|
| Interests & Tastes | .89 | .74 | .92 |
| Life Motto & Values | .97 | .18 | .71 |
| Work or Education | .92 | .79 | .93 |
| Relationships | .92 | .52 | .88 |
| Personality | .93 | .52 | .89 |
| Origin or Residence | .89 | .78 | .93 |
| Travel | .93 | .78 | .95 |
| Hospitality | .86 | .72 | .91 |

Table 4.3: Performance of sentence category classification.

We created a dataset of 5,248 profile sentences tagged with categories (we used the term “topics” in Step 1, but to avoid confusion with the term commonly used in the context of topic modeling in the NLP community, we refer to them as “categories” here).

Here we leverage the sentence level annotation dataset and trained eight binary classifiers to predict whether a sentence belongs to each category. We used the same pre-processing pipeline, and a one-gram bag-of-words model. We set the minimum threshold of token frequency to be 10, resulting in 616 features. We used a Bernoulli naive Bayes classifier, one for each category, and five-fold cross validation to evaluate the performance of the classification. The accuracy, F1-score and AUC for sentence category classification are listed in Table 4.3. The category *Life Motto & Values* has the worst F1 and AUC performance due to the extremely imbalanced label — there are very few sentences that were tagged to belong to this category.

We applied the trained classifiers on each sentence in the extended dataset, then adding up the classification results for sentences for the same profile into a vector of length eight representing how many sentences in the profile were tagged as belonging to one of the eight categories.

| Features | 6–19 words | 20–36 words | 37–58 words | 59–88 words | 89-179 words |
|------------------------|---------------|----------------|----------------|----------------|-----------------|
| WC | 57.5% | 53.8% | 46.9% | 43.9% | 50.0% |
| BOW | 60.8% | 58.5% | 62.6% | 59.4% | 58.8% |
| BOW + WC | 59.1% | 57.8% | 63.0% | 59.1% | 60.2% |
| LIWC + WC | 69.1% | 58.1% | 58.4% | 61.8% | 57.8% |
| Category + WC | 64.0% | 57.1% | 62.3% | 65.2% | 65.3% |
| Category + LIWC + WC | 65.9% | 59.4% | 65.9% | 65.9% | 68.0% |
| Best Model on Held-Out | 72.4% | 67.1% | 51.2% | 58.2% | 62.5% |

Table 4.4: Model performance (accuracy) summary by different length batch. Random baseline accuracy is 50%. Models compare profiles of similar lengths to predict relative perceived trustworthiness based on word count.

Models and Evaluation

We combine previously extracted features into different models and evaluate their classification performance using cross-validation. We chose the simple Word Count (WC) and BOW as baseline models. For LIWC and sentence category, we compared the performance of models using each set of features alone, and each plus Word Count, and found that Word Count improves performance; we only include the WC-enhanced models here (WC did not improve on BOW model, but BOW+WC is included here for completeness). We report the performance of all models in Table 4.4.

The bold numbers in Table 4.4 indicate the best performing model for each length batch. For the shortest profiles, the LIWC+WC combination achieves the best prediction result, while longer profiles benefit from including sentence category as features.

Evaluation on Held-Out Set

After picking the best performing model for each length batch, we re-fitted

the models on the entire cross-validation dataset to predict data from our completely disjoint held-out set. For the held-out set, we separated the profiles using the word count thresholds as defined in the training stage in to each batch, and obtained the perceived trustworthiness quartile tags using the thresholds obtained from training stage. We report the accuracy of prediction on held-out set in the last line of Table 4.4. Overall, the performance levels on the held-out set are comparable to other text-classification work [Danescu-Niculescu-Mizil et al., 2013b, Tan et al., 2014].

4.4.3 Factors Contributing to Perceived Trustworthiness

We conduct Lasso regression for profiles of different lengths to uncover factors that contribute to higher and lower perceived trustworthiness. We again use the top and bottom quartile of the data for each length batch in the cross-validation dataset for this analysis, using the *R* implementation of Lasso logistic regression (`cv.glmnet`) with default 10-fold cross-validation to choose the best parameter (λ) and using area under curve (AUC) as measure for goodness of fit. We report features that appear in more than 10% of the profiles as well as selected to be non-zero by Lasso in Table 4.5 and discuss the findings below.

4.5 Discussion

Sharing economy represents the third wave of the digitalization of social exchange. In this chapter, we took a look at how language cues establish trust in social exchange that sharing economy platforms are designed to unlock. The

work presented in this chapter makes several contributions to extend our understanding of trust in this new context. To begin with, a new coding scheme was introduced to categorize the language of self-disclose in Airbnb host profiles and to annotate the perceived trustworthiness of these profiles. We find that hosts use a variety of disclosure strategies, with some more successful than others, suggesting that platforms can better support users to convey trustworthiness by guiding what they disclose in profiles. In addition, computational framework has also been developed to distinguish between Airbnb host profiles of low and high perceived trustworthiness. We uncover features that are most predictive to higher perceived trustworthiness, such as *Hospitality* sentence category, and LIWC features *social* and *work*.

Specifically, in Step 1, the three studies reported make several contributions. First, we developed and validated a coding scheme for self-disclosure on the free-text portion of host profiles on Airbnb. The coding scheme describes eight topics that covers more than 90% of current discourse in host profiles. To the best of our knowledge, this is the first systematic coding scheme for analyzing self-disclosure in Airbnb profiles, or more generally, for profiles related to peer-to-peer sharing platforms.

The results of applying this coding scheme to the Airbnb profile dataset revealed that hosts most frequently write about *Origin or Residence*, *Work & Study* and *Interests & Tastes*. The least commonly disclosed topics were *Life Motto & Values*, *Relationships* and *Personality*. Study 1 in Step 1 also revealed that host type influenced the kinds of disclosures produced in host profiles, with on-site hosts revealing more information about their *Interests & Tastes* and *Personality* than remote hosts. These data are consistent with predictions from uncertainty

reduction theory [Berger and Calabrese, 1975], which predicted that on-site hosts will disclose more information about their *Interests & Tastes* and *Personality* to reduce potential guest uncertainty about whether they would enjoy interacting with an on-site host.

Our studies also drew on signaling theory [Spence, 2002] to understand what topics hosts disclose, and how guests perceive those disclosures. To assess the implications for signaling theory, it is informative to consider the results across both Study 1 in Step 1, which focused on the production of disclosures in host profiles, and Study 2 in Step 1, which examined how those disclosures affected the perceptions of trustworthiness. Signaling theory predicts that hosts will signal their trustworthiness by disclosing more assessment signals (e.g. *Origin or Residence, Work or Study*), which are more difficult to fake than conventional signals (e.g. *Life Motto & Values, Personality*) [Donath, 2007]. Signaling theory also predicts that, if hosts are optimizing their disclosures for trustworthiness, then guests should evaluate profiles with the most frequently observed topics as most trustworthy. The data from Studies 1 and 2 in Step 1 largely confirmed both of these hypotheses: hosts disclosed more assessment signals than conventional ones, and guest perceived profiles with more assessment signals as more trustworthy.

There were, however, some important exceptions to the theoretical predictions. Certain strategies, such as demonstrating *Hospitality* or sharing one's *Interests & Tastes*, proved to be more successful than expected for conventional signals while other strategies, such as providing only one's *Origin & Residence*, proved less successful. The language of *Hospitality* is more successful potentially because it provides more information about expected interactions, therefore re-

ducing uncertainty in future exchange and increasing trust. Providing a welcome or greeting, or providing reasons for hosting, alone but preferably in combination with more assessment signal disclosures (e.g., *Hospitality* combined with *Origin & Residence*), was an important strategy that had a strong and positive effect on perceptions of trustworthiness. These findings suggest that signaling with conventional signals but that provide information about one's hospitality or interests can enhance trustworthiness in Airbnb profiles.

Finally, we demonstrated that perceived trustworthiness matters for decision-making in this context. The perception of trustworthiness from Study 2 in Step 1 predicted participants' decisions in a forced choice experimental task in Study 3 in Step 1, especially for profiles that are relatively short (less than 20 words). We also showed that when profiles are short, perceived trustworthiness almost perfectly predicts choice, whereas when the profile length increases, other factors appear to also influence choice. This may suggest a nuanced role of trust in decision-making — there is a threshold of trust that is needed to pass muster, but other factors (e.g. homophily [McPherson et al., 2001]) may weigh in once trustworthiness is no longer the issue.

This research suggests that the Profile as Promise framework [Ellison et al., 2012] is a useful approach for understanding how hosts and guests produce and evaluate disclosures in Airbnb profiles. Hosts disclosed information about themselves that they perceived as relevant and of interest to potential guests, and their promises were evaluated based on their trustworthiness, as predicted by signaling theory and uncertainty reduction theory. This study suggests that the concept of a promise, or psychological contract, can be usefully applied beyond online dating profiles [Ellison and Hancock, 2013] and résumés [Guillory and

Hancock, 2012] to peer-to-peer sharing platforms such as Airbnb.

At the same time, in Step 2, the computational approaches allow us to gain more detailed insights on what factors and language features contribute to higher perceived trustworthiness. The key factors contributing to higher perceived trustworthiness are *Hospitality* sentence category, and LIWC features *social* and *work*. The sentence category *Hospitality* contributes to higher perceived trustworthiness for profiles longer than 37 words. The effectiveness of hospitable language strengthens the findings of [Ma et al., 2017a]. The *social* LIWC category also contributes to higher perceived trustworthiness, potentially through the mechanism of uncertainty reduction [Berger and Calabrese, 1975]. Providing information about one's social relationships can make hosts appear more "real". Finally, LIWC feature *work* predicts higher perceived trustworthiness for all profiles shorter than 59 words, potentially also through the mechanism of uncertainty reduction.

In contrast, LIWC feature *leisure* and sentence category *Interest & Tastes* are contributing *negatively* to perceived trustworthiness for profiles between 6–19 words and 37–58 words respectively. The negative effect of these features may suggest a separation between the need of sociability and the ability to provide standard goods and services in sharing economy.

Comparing these features that were found to be significant in our work with previous research, we uncover a potential discrepancy between the language that is *perceived* to be trustworthy, and the *actual* trustworthiness of individuals. In terms of perception, as we see in our work, and previous work on crowd funding, LIWC features *social* and *work* contribute to higher perceived trustworthiness or higher likelihood of a project being funded [Mitra and Gilbert, 2014]. However,

in terms of actual trustworthiness, *social* language is found to be associated with higher loan default rates [Netzer et al., 2016]; and in online dating profiles, online daters used more *work* related words [Toma and Hancock, 2012] when their photos are less accurate. Expanding on the discrepancy between perceived and actual trustworthiness would be important future work.

4.5.1 Design Implications

Our findings have direct implications for improving the design of profile pages on sharing economy platforms, with the view of encouraging trustworthiness and improving the rate of transactions. Our results suggest that hosts should be encouraged to disclose more information, and that this information should come from a diverse set of the eight categories identified in the coding scheme from our study. With knowledge of the profile features that may promote trust, interfaces for creating and editing profile text could encourage individuals to write more, and focus on the key categories exposed above. Automatic text analysis mechanisms could also be used to classify text into categories, and suggest other topics to improve breadth and ultimately perceived trustworthiness.

4.5.2 Limitations

There are some important limitations to this work. First and foremost, we opted to prioritize our theoretical understanding of trustworthiness in profiles, over developing an ecologically valid measure of the profile text's effect on host choice. As mentioned above, host choice on Airbnb can be impacted by many factors,

including (most trivially) the price and characteristics of the rental property. Nevertheless, the experiment we ran on host choice allowed us to make causal claims regarding the profile text's impact on guest decision-making.

A related limitation of the work is the fact that it ignores dyadic and dynamic determinants of trustworthiness. A key mechanism of uncertainty reduction theory involves dyadic reciprocity and exchange [Berger and Calabrese, 1975]. In this work, we only examined a single-sided, one-time disclosure by hosts. It would be important to consider the effect of the dyadic properties of hosts and guest, and how they relate to trustworthiness and trust. Understanding how impressions of perceived trustworthiness form and evolve through conversations between hosts and guests would be another complex and interesting problem to tackle.

Our dataset only includes U.S. large cities. As a result, the findings may not generalize to hosts in smaller cities, though nothing in our findings would necessarily suggest that this would be the case. We did not consider gender and cultural differences in this work, either. In a preliminary investigation, we inferred the gender of hosts from their first names, but did not find significant differences in disclosures between hosts of different gender. Future work can dive deeper into patterns of self-disclosure by individuals of different gender and cultures, potentially helping to combat discrimination or potential biases known to exist on sharing economy platforms [Thebault-Spieker et al., 2015].

While this work uses quantitative approach, another approach would have used other qualitative methods, such as interviews with hosts about their profile construction strategies. For example, how do Airbnb hosts present a trustworthy facade while balancing other important aspects (e.g. privacy)? Other research

has qualitatively examined the experiences of hosts [Lampinen and Cheshire, 2016] but to date has not considered profile construction work.

A key question that requires future work is whether our findings are unique to the context of lodging in sharing economy, more specifically to Airbnb, or maybe they apply more generally in sharing economy platforms. While we believe some features we identified, for example some LIWC features in the different models, apply more generally, other features are context specific. For example, the sentence category *Hospitality* is specific to the lodging context, though at the same time a version of it can transfer to other domains (e.g., promise of service). Future work can expand our results to other instances of sharing economy platforms.

Finally, because we relied on crowd workers to provide ratings of perceived trustworthiness, our annotation and the resulted algorithms trained on the annotated data run the risk of having potential bias. For example, it is possible that people have stereotypes towards certain professions, such as artists. When that is revealed in self-disclosure in profiles, it is possible that such stereotypes will be penalized in terms of perceived trust. Another potential bias is that people might perceive women hosts as more trustworthy. Future work can look into whether systematic bias exist in the annotation.

4.6 Conclusion and Extensions

In this chapter, we conducted a two-step inquiry of how language establishes trust through self-disclosure in the context of online sharing economy platform — Airbnb. Through the computational modeling of Airbnb host profiles, we show

our frameworks can distinguish between Airbnb host profiles of low and high perceived trustworthiness. In addition, we uncover language cues that are most predictive to higher perceived trustworthiness, including cues of hospitality, cues related to social context, and cues related to work. This shows that in addition to uncertainty reduction, signaling plays a very important role in self-disclosure to establish trust through language on sharing economy platforms.

Taken together, the last and this chapter represent the first focus of networked trust: cues in Computer-Mediated Communication, including image cues and language cues. As shown in both chapters, we can develop custom algorithms to predict high or low quality images, or to predict profiles that are perceived as high or low in trustworthiness. These algorithms can be fitted to provide feedback for users to take better images, write better profiles, or even generate profiles altogether, to optimize for trust.

Such potential for algorithmic augmentation of online presentation raises new questions about the reliability of cues in Computer-Mediated Communication. As a result of algorithmic mediation, what used to be *Computer-Mediated Communication* (CMC) is turning into **AI-Mediated Communication (AI-MC)**: interpersonal communication not simply transmitted by technology but augmented — or even generated — by algorithms to achieve specific communicative or relational outcomes. In AI-Mediated Communication, an AI system operates on behalf of the communicating person, e.g., by augmenting, generating or suggesting content. AI-Mediated Communication is distinct from traditional CMC technologies that primarily transmit messages, and from typical machine-authored texts that do not represent a person.

In one of the follow-up works to the study presented in this chapter, we

observed the effects of AI-Mediated Communication in the first attempt to conceptualize AI-Mediated Communication [Jakesch et al., 2019]. Through a series of three online experiments, we examine how the belief that a computer system has generated a host's profile changes whether the host is seen as trustworthy by others. We observed that (1) when people are presented with *all AI-generated* profiles they trust them just as they would trust *all human-written* profiles; (2) when people are presented with a *mixed* set of AI- and human-written profiles, they mistrust hosts whose profiles they believe were generated by AI.

As algorithms continue to gain in importance in mediating our communication online, it is important to understand the theoretical and practical implications of algorithmic mediation, especially on interpersonal trust. Future studies can continue to investigate AI-Mediated Communication's impact on interpersonal trust as new applications continue to appear on the market.

| Profile Length | Positive Features | | Negative Features | |
|----------------|---|--|---------------------------------|---|
| | Sentence Category | LIWC | Sentence Category | LIWC |
| 6–19 words | (Not included) | readability, comma, article, conjunction, social, affect, causation, health, work, achieve | (Not included) | word per sentence, exclamation mark, present tense verb, adverb, quantifier, human, tentative, perceptual process, sexual, relativity, motion, space, leisure |
| 20–36 words | Travel | we, social, positive emotion, cognitive process, sexual, work, home | — | parenthesis, adverb, prepositions, perceptual process |
| 37–58 words | Work or Education, Relationships, Hospitality | parenthesis, we, article, auxiliary verb, past tense verb, social, family, friend, certain, work | Interests & Tastes, Personality | comma, dash, exclamation mark, period, negation, quantifier, insight, motion, leisure |
| 59–88 words | Relationships, Hospitality | we, social | — | — |
| 89–179 words | Hospitality | social, inclusive | — | — |

Table 4.5: Factors contributing to higher and lower perceived trustworthiness by different length batch. LIWC categories “we” and “social” are important positive indicators, potentially through social warranting. The LIWC “sexual” category contained mostly the word “love”.

CHAPTER 5

NETWORKS OF TRUST: WHEN DO PEOPLE TRUST THEIR SOCIAL GROUPS?

5.1 Introduction

Last two chapters covered the first focus of networked trust — cues in Computer-Mediated Communication. This chapter presents a case study exploring the second focus of networked trust: trust in social exchange that are embedded in social *networks*. In particular, I study how **social networks contribute to trust in the context of Facebook groups**.¹ Again here we use a survey-based measure to assess the perceived trustworthiness of other members in a social group.

As one of the most prominent social network sites (SNSs), Facebook is among the second wave of the digitalization of exchange — the digitization of social relationships. SNSs enable users to articulate and make visible their social networks [Boyd and Ellison, 2007], and they require that people trust each other with their personal information. SNSs brought new challenges about trust, especially in relation to privacy [Marwick and Boyd, 2011], but they also present new opportunities to understand how social structure contributes to trust. For example, people’s friendship network structure has been shown to be informative for predicting romantic relationships [Backstrom and Kleinberg, 2014].

This chapter leverages the opportunities to study **trust in social groups** on SNSs. Social groups are important social structures through which communities

¹This work was published at CHI 2019 as *When Do People Trust Their Social Groups?* [Ma et al., 2019a]

are formed. On Facebook, billions of people engage with social groups every month. Trust is attributed to contribute to the success of social groups by encouraging people to interpret others' actions and intentions favorably, thereby facilitating cooperation and a sense of community [Gambetta, 1988, Misztal, 2013, Uzzi, 1996, Bachmann, 2001, Dirks, 1999, Preece and Maloney-Krichmar, 2005]. In groups, trust increases member satisfaction and task performance [Walther and Bunz, 2005], reduces conflict [Gambetta, 1988, Walther and Bunz, 2005], and promotes effective response to crisis [Meyerson et al., 1996]. Previous research has examined how different factors such as size [Brewer, 1991, Denters, 2002, Zelmer, 2003], group cohesiveness [Hogg, 1993], and activity [Walther and Bunz, 2005] may impact people's trust in their social groups, both online [Holtz et al., 2017] and offline [Rotter, 1971]. However, previous studies tend to be small in scale, limited to specific contexts (e.g., online marketplaces), or only consider a specific type of group (e.g., organizations [Mayer et al., 1995, Colquitt et al., 2007]). Studies that address these three limitations may enrich our understanding of how trust is formed in social groups more generally. In particular, as people in a group are connected in different ways, social network structure of the group may play an important role in structurally determining how much people trust the group.

In this work, we build on rich prior literature on trust to present a framework for predicting an individual's trust in a social group, and examine how differences at the individual and group levels predict that trust, especially the social network structure of groups. We focus our analysis on Facebook Groups², a Facebook feature that "allows people to come together to communicate about shared interests" [Facebook, 2018]. As of May 2018, 1.4 billion people use Face-

²We use "Groups" to refer to the Facebook product, and "groups" to refer to actual social groups on Facebook.

book Groups every month [Perez, 2018]. By combining a survey ($N=6,383$ valid responses) of individuals using Facebook Groups with aggregated behavioral logs, we are able to investigate, across a diverse sample, how an individual's trust in a group relates to characteristics of the individual, the group, and the individual's membership in that group.

The survey asked individuals about their general attitudes towards others and trust towards a Facebook group that they were a member of. While prior work has shown that an individual's general propensity to trust others influences their trust in a particular group [Boss, 1978, Butler Jr, 1999, Ridings et al., 2002, Ferguson and Peterson, 2015], we additionally examine the role of other individual-level differences (e.g, general attitudes towards risk-taking).

We combine these survey results with aggregated behavioral and descriptive data on Facebook Groups. This allows us to study the role of five categories of features that characterize either the group or the respondent's relationship with the group, based on prior literature: (1) basic properties of the group (e.g., size, membership privacy policy) [Kraut and Fiore, 2014]; (2) group category [Denson et al., 2006]; (3) group activity [Kraut and Fiore, 2014]; (4) group homogeneity [Moser et al., 2017]; and (5) the friendship-network structure of the group [Holtz et al., 2017].

We find that these variables robustly predict participants' trust in a particular group, with both individual and group characteristics contributing predictive value (adjusted $R^2=0.26$). In particular, an individual's trust in a group was most strongly predicted by their general perceived social support, the group's average clustering coefficient, and their degree centrality in the group. We also show that trust in a group can be estimated using only observational data.

While these results support previous findings showing that intragroup trust decreases with increasing group size and increases with membership restriction [Zelmer, 2003, Denters, 2002, Brewer, 1991, La Macchia et al., 2016, Moser et al., 2017], we find that these trends only hold up to a certain point. When the size of a group exceeds 150 members (roughly Dunbar's number, or the expected cognitive limit beyond which social relationships are difficult to maintain [Dunbar, 1992]), the membership policy of the group (public v.s. private) ceases to play a predictive role. Moreover, in deciding how much to trust a group, we show that group size matters less to individuals with a higher general propensity to trust.

Further, previous work suggests that people trust groups in which they are more active [Cartwright and Zander, 1953], but we find that only certain types of activities are associated with trust: people "like" and "comment" more in groups they trust but do not necessarily post more, suggesting that trust is associated more with directed communication than with information sharing.

Finally, we show that trust in groups is associated with both individual- and group-level outcomes. Increased trust leads to individuals being more likely to form friendships with other members of the group, but is also associated with the group being less likely to grow larger in size.

In summary, we (1) present results of a large survey of individuals' trust attitudes towards their social groups (6,383 unique groups); (2) examine how characteristics of both the individual and group contribute to trust in a group; and (3) show how this trust affects future individual- and group-level outcomes. A deeper understanding of how these factors collectively contribute to trust in groups can better equip communities to foster trust among their members.

5.2 Related Work

5.2.1 Determinants of Trust in Groups

What contributes to trust in groups? Here we review relevant literature that guide the selection of our feature sets in predicting trust in groups.

Individual Differences

Trust in groups can be mediated by one's disposition to trust others, as it correlates with one's initial intentions to trust a group, especially in ambiguous situations [Gill et al., 2005]. A disposition to trust can positively impact trust in different settings, including trust between individuals [Yakovleva et al., 2010], in communities [Ridings et al., 2002], in organizations [Kantsperger and Kunz, 2010], or in online services [Wu et al., 2010]. Similarly, a disposition to trust increases trustworthiness evaluations given to Airbnb hosts [Ma et al., 2017a], though in other settings, a disposition to trust was not associated with trust in peer sellers [Jones and Leonard, 2008].

Past work also suggests an inverse relationship between risk aversion and trust [Abraham et al., 2017] — the more comfortable an individual is with taking risks, the higher the trust they have in groups.

Further, prior literature treats membership of voluntary associations as an indicator of trust [Putnam, 2000, Putnam, 1993]. Thus, greater in-group loyalty, as well as perceived social support from others, should both be linked with higher trust in groups due to increased group participation and perceived social capital.

Group Characteristics

Trust in groups may also stem from basic properties of the group such as its size [Brewer, 1991, Denters, 2002, Zelmer, 2003]. For instance, experiments have shown that people identify more strongly with smaller groups [Simon and Brown, 1987]. In addition, groups that have existed for longer periods of time should also be trusted more, as they have more time to develop established norms and culture that are beneficial for group trust. Past research has also described how secrecy can build community [Fine and Holyfield, 1996] and shown that group cohesiveness promotes trust [Stokes, 1983]. Recent qualitative work on Facebook Groups also suggests that by making membership exclusive and screening new members, trust can be enhanced [Moser et al., 2017].

Homogeneity, which relates to cohesiveness, may also contribute to trust. People tend to be closer to and trust others who are similar to them [McPherson et al., 2001]. Research has also found a relationship between gender and age homophily and increased trust [Ahmad et al., 2011, Abrahao et al., 2017].

Higher levels of group activity are also linked with greater trust [Cartwright and Zander, 1953, Walther and Bunz, 2005]. Increased social interaction provides “opportunities for people to get acquainted, to become familiar with one another, and to build trust” [Ren et al., 2007], thus leading to higher familiarity, and in turn, greater trust [Gulati, 1995].

Network Characteristics

Beyond group characteristics mentioned above, the overall structure of relationships between individuals in the group, as well as the individual’s embeddedness

the group’s social network may mediate trust. A person’s number of friends and the connections among these friends can both increase the likelihood of them joining a community [Backstrom et al., 2006, Ugander et al., 2012]. As dense networks promote cooperation and social norms, they are also likely to be associated with increased trust [Coleman, 1988]. In buy-and-sell groups on Facebook, network density and the degree centrality of the seller are positively correlated with an intention to transact, which may signal higher trust in the group [Holtz et al., 2017]. Our work uses similar features but directly measures trust via a survey, and considers the role of network features within a much large set of variables.

This rich prior literature motivates our analysis in this work, in which we conduct a large-scale survey and analyze behavioral data to show how individual- and group-level differences help predict trust in groups. Our research questions are as follows: (a) Can a baseline model that accounts only for individual attitudes predict trust in groups? (b) What is the relative contribution of the different sets of group-level features (basic group properties, group category, activity, homogeneity, and structural properties) on trust in groups beyond the baseline model?

5.3 Methods

In this work, we conducted a survey of 10,000 respondents to a random sample of active Facebook Groups users in the U.S. People were invited to participate in the survey via an ad on Facebook. The survey was designed to measure both individual attitudes as well as trust in one of the randomly selected Facebook

groups of which they were members. We augmented this survey data with self-reported demographic data such as age and gender and server logs of these individuals' activity and friendships on Facebook. All log data was de-identified and analyzed in aggregate on Facebook's servers; researchers did not view any identifiable data nor access any specific posts in any groups. The study was approved by an internal Facebook board as well as Cornell's Institutional Review Board under protocol #1805008006.

5.3.1 Sampling

The survey was issued to unique individual-group pairs. We used the following sampling strategy to identify eligible survey candidates. First, we identified Facebook groups that had at least five members and that had a majority of their members located in the U.S. We then identified people in the U.S. who belong to at least one of these groups, and that had at least one interaction (e.g., creating, liking, or commenting on a group post) in the past 28 days. We then sampled eligible individual-group pairs, de-duplicating by both individual and group at random. The sampling was also stratified by group size (the number of members in the group) to better capture behavior across both smaller and larger groups. We note the following bias introduced by our sampling method: compared to all individuals who actively used groups in the past 28 days, our participants tended to be 8.7% older and were 17.5% more likely to be women.

5.3.2 Survey Design

The survey consisted of two sections: a section on individual differences regarding the participant's general attitudes towards others, including disposition to trust and related concepts; and a section on trust in a specific Facebook group. Each section had four items, shown in Table 5.1. The order of questions was randomized within a section. Participants were asked to report the extent to which they agreed or disagreed with each statement on a five-point Likert scale.

For the section on general attitudes towards others, we measured disposition to trust through an adaptation of the generalized trust question in the World Values Survey [WVS, 2018]. The original question elicits a dichotomous response, worded as: "Generally speaking, would you say that most people can be trusted or that you need to be very careful in dealing with people?" We instead used a more granular five-point Likert scale, which has been shown to be more reliable [Miller and Mitamura, 2003]. We also included measures of concepts related to disposition to trust reported in previous literature, including general social support [Barrera Jr and Ainlay, 1983, Vigoda-Gadot and Talmud, 2010, Hether et al., 2014], in-group loyalty [Van Vugt and Hart, 2004], and risk aversion [Miller and Mitamura, 2003].

To measure an individual's level of trust in a Facebook group of which they were a member, we created a four-item scale to measure trust in groups (section two in Table 5.1), based on previous literature. This scale is based on the framework of ability, integrity, and benevolence by Mayer et al. [Mayer et al., 1995] and Schoorman et al. [Schoorman et al., 2007], and also adapts measures from several interpersonal trust scales including Rotter Interpersonal Trust Scale [Rotter, 1971], the Specific Interpersonal Trust Scale [Johnson-George and Swap, 1982], and a

newer “predisposition to trust” scale [Ashleigh et al., 2012].

In addition, to better understand what people use the group for, we also asked participants to use the taxonomy below to describe the group category:

- Friends & Family: e.g., close friends, extended family
- Education & Work: e.g., college, job, professional
- Interest-Based: e.g., hobby, book club, sports
- Identity-Based: e.g., lifestyle, health, faith, parenting
- Location-Based: e.g., neighborhood or local organization
- Other

These categories were identified in previous qualitative research, where we surveyed people who used Facebook Groups and asked them to describe a group they were part of (e.g., “my family”). In our work, participants were requested to select all categories that applied to the group they were surveyed on, and we treated each group category as a binary variable. In our sample, around 34% of the groups were tagged as interest-based groups (most common), followed by 20% friends & family groups. The first five categories capture most of the groups (covering 89%).

5.3.3 Data and Statistical Approaches

In addition to data from the survey, we examined properties of groups including their sizes and membership privacy policies. For each group, we also looked at an individual’s activity in the group (e.g., time spent, likes, comments, and posts

| General Attitudes Towards Others | |
|---|--|
| Disposition to trust | Most people can be trusted. |
| General social support | There are people in my life who give me support and encouragement. |
| General risk attitude | I'm willing to take risks. |
| General in-group loyalty | I would describe myself as a "team player". |
| Trust in a Group | |
| Care | Other members of the group care about my well-being. |
| Reliability | Other members of this group can be relied upon to do what they say they will do. |
| Integrity | Other members of this group are honest. |
| Risk-taking | I feel comfortable sharing my thoughts in this group. |

Table 5.1: Trust in groups survey items. Participants reported the degree to which they agreed or disagreed to each of the survey items on a five-point Likert scale.

made), the group's overall activity, as well as group members' friendships with each other.

Out of the 10,000 survey responses we received, we filtered responses based on the completeness of the survey, as well as availability of self-reported and log data. In the end, we retained 6,383 responses for our main analysis.

The main statistical techniques we used were multiple linear regression, random forests, and logistic regression. We first built a baseline model predicting trust in groups using variables capturing individual-level differences. Then, we identified five different sets of group-level features, conducted analysis on how much each set of feature improved the baseline model, and interpreted the relationship between each feature and trust separately. We next combined all features in a random forest model and compared the importance of each set of features in the combined model. Finally, we used logistic regression to predict group outcomes such as the densification of the friendship network within the

| Variable | Mean | SD | Correlations | | |
|---|------|------|--------------|---------|---------|
| | | | 1 | 2 | 3 |
| General Attitudes Towards Others | | | | | |
| 1. Disposition to trust  | 3.33 | 1.07 | | | |
| 2. General risk attitude  | 3.68 | 0.99 | 0.18*** | | |
| 3. General in-group loyalty  | 4.54 | 0.80 | 0.23*** | 0.16*** | |
| 4. General social support  | 4.34 | 0.84 | 0.24*** | 0.18*** | 0.36*** |
| Trust in Groups | | | | | |
| 1. Care  | 3.90 | 1.08 | | | |
| 2. Reliability  | 4.05 | 0.98 | 0.62*** | | |
| 3. Integrity  | 4.20 | 0.95 | 0.60*** | 0.67*** | |
| 4. Risk-taking  | 4.09 | 1.06 | 0.59*** | 0.54*** | 0.56*** |

Note: * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

Table 5.2: Descriptive summary of survey measures, including general attitudes and trust in groups. Sparklines represent the histogram of each measure. (N=6,383)

group. When appropriate, we log-transformed the data (e.g., group size) and note the transformation when reporting coefficients.

5.4 Results

Trust in groups was measured in our survey across four dimensions: care, reliability, integrity, and risk taking. As shown in Table 5.2, these dimensions of trust in groups are highly correlated ($\rho \geq 0.54$; Cronbach's $\alpha = 0.86$). Thus, we defined a composite "trust in groups" score as the mean of all four dimensions and report findings with respect to this composite score.

| | <i>Dependent variable:</i> | | | | |
|-------------------------|---------------------------------|----------------------|---------------------|--------------------|--------------------|
| | Trust in groups composite score | | | | |
| | (1) | (2) | (3) | (4) | (5) |
| (Intercept) | 4.07*** (0.04) | 3.51*** (0.05) | 3.18*** (0.06) | 2.46*** (0.07) | 1.98*** (0.08) |
| Age | -0.001* (0.001) | -0.003*** (0.001) | -0.002** (0.001) | -0.001* (0.001) | -0.001* (0.001) |
| Female | 0.09*** (0.02) | 0.06** (0.02) | 0.08*** (0.02) | 0.05* (0.02) | 0.02 (0.02) |
| Disposition to trust | | 0.19*** (0.01) | 0.17*** (0.01) | 0.13*** (0.01) | 0.11*** (0.01) |
| Risk attitude | | | 0.09*** (0.01) | 0.07*** (0.01) | 0.05*** (0.01) |
| In-group loyalty | | | | 0.21*** (0.01) | 0.16*** (0.01) |
| Social support | | | | | 0.19*** (0.01) |
| Adjusted R ² | 0.003 | 0.06 | 0.07 | 0.11 | 0.14 |

Note: *p<.05; **p<.01; ***p<.001

Table 5.3: Baseline model predicting trust in groups using demographics, disposition to trust, risk attitude, in-group loyalty, and social support. (N=6,323 after removing missing age and gender observations)

5.4.1 Individual Differences and Trust

We start by predicting trust in groups using individual attitudes as well as demographic information (see in Table 5.3), which prior literature has associated with differences in one’s disposition to trust [Taylor et al., 2007].

Demographics

We found that demographic factors such as the age and gender of participants capture almost no variance of trust in groups (see Model 1 in Table 5.3). This result partially contrasts with the prior work that found a relationship between

these demographic factors and one's disposition to trust [Taylor et al., 2007]. To better understand this result, we tested a model that used demographic variables to predict participants' disposition to trust rather than trust in groups. While we found that older people were more trusting than young people ($\beta=0.006, p<.001$) and women were more trusting than men in general ($\beta=0.12, p<.001$), very little variance in disposition to trust is explained by these demographic factors [$R^2=0.01, F(2, 7174)=36.1, p<.001$]. In other words, demographic characteristics explain neither an individual's disposition to trust nor their trust in groups.

General Attitudes Towards Others

How do an individual's general attitudes towards others predict their trust in groups? Corroborating prior work, one's general disposition to trust significantly predicts one's trust in groups (see Model 2 in Table 5.3). However, other factors also play significant roles (Models 3–5 in Table 5.3). Notably, the individual's perceived social support ($\beta=0.19, p<.001$) and their general stated in-group loyalty ($\beta=0.16, p<.001$) contributed more to the prediction of trust in group than one's disposition to trust ($\beta=0.11, p<.001$). A willingness to take risks ($\beta=0.05, p<.001$) was least predictive. Altogether, these factors capture a significant amount of the variance in trust in groups (adjusted $R^2=0.14$).

5.4.2 Group Differences and Trust

To understand the relationship between group characteristics and trust in groups, we identified five distinct sets of group-level features (see Table 5.4). In this section, we measure the incremental predictive value of each of these sets of

| Feature Set | Features |
|----------------------|---|
| Basic Properties (5) | Group size, privacy type, group tenure, number of admins/moderators |
| Category (5) | Self-reported group category |
| Activity (6) | Group-level and participant-group-pair level time spent, number of posts, number of likes or comments |
| Homogeneity (3) | Diversity of group age, gender, and similarity between participant and group average |
| Structural (5) | Network density, average clustering coefficient, participant degree centrality, cliquishness of participant's friends in the group, average number of mutual friends with group members |

Table 5.4: Five sets of group-level features used for predicting trust in groups.

group-level features, after controlling for the individual differences discussed above. Here, we use “baseline model” to refer to a model that only includes the individual differences (Model 5 in Table 5.3). For each feature set, we add the features as independent variables in the multiple linear regression model to the baseline model. In each subsection, we report how much the model gains from the additional features. We validated that the coefficients of the individual differences features in the baseline do not change significantly when we include each new feature set.

Basic Group Properties

The first set of group-level features consisted of group size, privacy type, group tenure (how long a group has existed), the number of group admins, and its number of moderators. Adding these features to the baseline model increased the model's adjusted R^2 by 0.08 ($p < .001$). The most significant predictor of trust was group size. Consistent with previous work on trust and group sizes [Brewer,

1991, Denters, 2002, Zelmer, 2003], people had lower trust in bigger groups ($\beta = -0.15$ on log scale, $p < .001$).

Apart from a group's size, a group's privacy type can also affect perceptions of trust. On Facebook, group admins can set the group to be "public", "closed", or "secret". Public groups are accessible to non-members, while closed and secret groups are only accessible to current members; closed groups differ from secret groups in whether their existence is known to non-members. We found no significant differences between closed and secret groups, so we analyzed them together as "private" groups.

Controlling for group size (public groups are 68% larger than private groups), we found that people trusted public groups *less* than private groups ($\beta = -0.07$, $p < .01$), as suggested in prior work [Moser et al., 2017].

Notably, we found an interaction effect between group size and privacy type in predicting trust ($\beta = 0.04$, $p < .01$): the larger the group, the smaller the difference there is between trust in private and public groups. To see how quickly this difference between group types dissipates, we conducted a series of t-tests in which we compared the mean difference in the trust composite score between public and private groups above a certain size threshold, starting from 10 in increments of 10. These tests show significant differences between groups larger than the threshold until the threshold exceeds 150, where we no longer observe a significant difference between public and private groups ($p > .05$).

Incidentally, this size threshold roughly corresponds to Dunbar's number — the maximum number of stable social relationships a person can maintain due to limitations in cognitive resources [Dunbar, 1992]. Smaller private groups

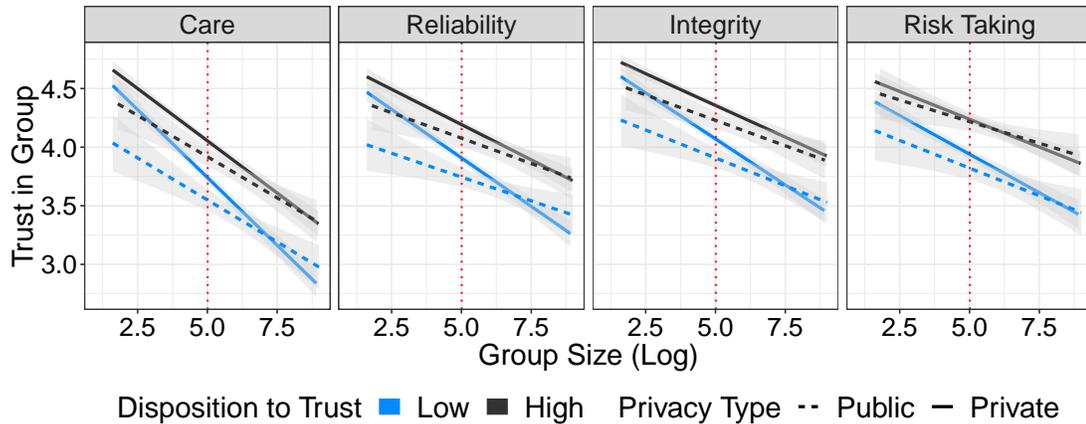


Figure 5.1: The relationship between trust in groups and group size, for each dimension (panels), across groups of different privacy types (line style) and individuals with different propensity to trust (line color). Dunbar’s number (150) is marked by a vertical red dotted line.)

provide control and exclusivity over membership, thus allowing members to foster a shared sense of identity [Moser et al., 2017]. Once the group becomes too big, that shared identity might be lost, resulting in no difference between large groups that are public or private.

Figure 5.1 summarizes the impact of group size on trust in public and private groups, as well as the effect of an individual’s disposition to trust (we consider composite scores >3 to be high and ≤ 3 to be low). The figure shows that having a high disposition to trust (black lines) and a group being private (solid lines) are both factors that contribute to trust in groups. But while the effect of privacy decreases with size (dashed and solid lines cross), the reverse is true for an individual’s disposition to trust. An interaction effect between group size and individual’s disposition to trust ($\beta=0.01, p<.01$) shows that people with a greater disposition to trust others were less sensitive to changes in group size (visually represented by gentler slope of black lines compared to blue ones in Figure 5.1).

Other basic group properties also relate to trust. Longer group tenure predicts higher trust ($\beta=0.04$ on log scale, $p<.001$), potentially because older members have more stable group relationships and are more familiar with other group members [Walther and Bunz, 2005]. The number of admins also predicts higher trust ($\beta=0.10$ on log scale, $p<.001$). This finding is consistent with previous work that found that groups with more admins tended to survive longer than groups with fewer admins [Kraut and Fiore, 2014]. The number of moderators is a much weaker predictor of group trust.

Group Category

As previously described, participants in our survey labeled groups as belonging to one or more of six categories. Including group category as multiple binary variables to the baseline model significantly improved trust predictions ($p<.001$), increasing the model's adjusted R^2 by 0.05. To illustrate differences in trust across these categories, we also conducted an ANOVA and plotted the average trust in groups by category in Figure 5.2. Groups marked as "other" were excluded from this analysis. Post-hoc Tukey tests showed that people trust friends & family groups the most, followed by identity-based and education & work groups ($p<.001$). They trust interest- and location-based groups least ($p<.001$).

Why does trust differ by group category? For friends & family groups, high trust is a strong sign of bonding social capital [Putnam, 2000]. Identity-based groups (e.g., parenting groups) and education & work groups elicit trust by establishing a shared identity among group members [Moser et al., 2017]. Finally, interest- and location-based groups may represent less bonding and more bridging social capital [Granovetter, 1985], especially for information sharing.

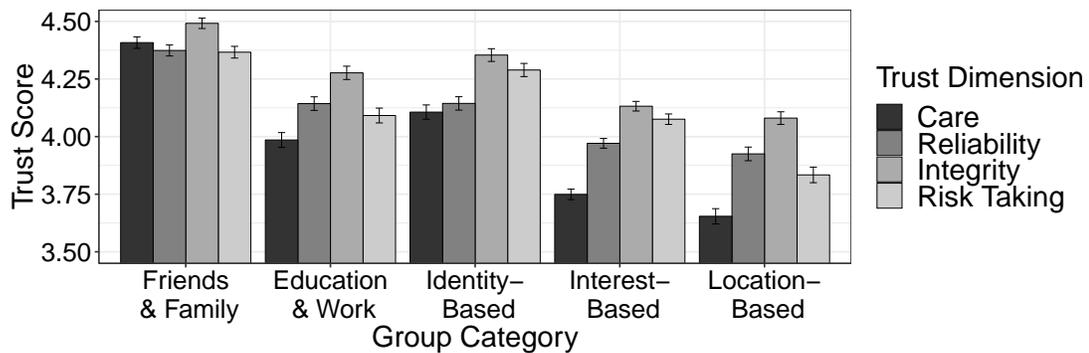


Figure 5.2: People have the highest trust in friends and family groups, and lowest in interest- and location-based groups.

People use these groups more as places to transact and exchange (both physical goods and information) than as places to build relationships [Granovetter, 1985]. By comparing groups across different categories, we can develop a more holistic understanding of trust across different types of social groups that also draws on insights from previously isolated studies [Moser et al., 2017, Holtz et al., 2017].

Activity

Here, we studied both a survey participant’s activity in a group as well as the overall group activity across all members. Measures of activity include time spent in the group and the number of actions (posts, likes, or comments) taken in the group, averaged across the 28 days preceding the survey. In the case of public groups, activity also included contributions from nonmembers. An individual’s overall site engagement was not predictive of trust, and thus was excluded from our analyses. Including activity features (time spent, group activity, and participant in group activity) to the baseline model improves its adjusted R^2 by 0.04 ($p < .001$).

As many activity features are correlated, we report coefficients when the feature is independently added to the baseline model. Time spent in the group, both by the individual ($\beta=0.04$, $p<.001$) and by other group members ($\beta=0.05$, $p<.001$) independently predicts higher trust in groups. Overall, the number of posts per member ($\beta=0.07$, $p<.001$), and the number of likes and comments per post ($\beta=0.07$, $p<.001$) were also both independently associated with higher trust. However, the number of comments and likes a participant made in a group was associated with higher trust ($\beta=0.10$ on log scale, $p<.001$), but not the number of posts the participant wrote.

Why is this the case? Posting in a group may be influenced by a variety of factors other than trust (e.g., self-esteem [Forest and Wood, 2012]). In contrast, likes and comments are forms of directed communication that people use to maintain relationships with others [Ellison et al., 2014] and may therefore be more conducive to building trust.

Homogeneity and Homophily

Trust may also be influenced by how similar people in a group are to each other (homogeneity), and how similar an individual is to others in the group (homophily).

For each group, we measured age and gender diversity by computing the gender entropy of the group's members and the standard deviation of their ages. To measure homophily, we constructed a simple distance measure based on the approach of [Abraham et al., 2017]. If the participant had the same gender with the majority of the group members, we coded the gender distance as 0, otherwise

1. If the participant's age was within 5 years of the average age of the group, we coded the age distance as 0, otherwise 1. The total distance from average group members was calculated as the L_1 distance, i.e., the sum of gender and age distance $\in (0, 1, 2)$. As different types of groups may have different demographic compositions, we controlled for group category in this analysis.

Adding homogeneity and homophily features to the baseline model results in a small improvement (increased adjusted R^2 by less than 0.01, $p < .001$). Nonetheless, we found that both gender ($\beta = 0.04$, $p < .001$) and age homogeneity ($\beta = 0.04$, $p < .001$) were associated with higher trust.

Surprisingly, homophily, measured as described above, was not predictive of trust in groups. This contrasts with findings in previous work on trust and homophily in dyadic exchange, which found that trust increases with gender and age homophily [Abraham et al., 2017, Ahmad et al., 2011]. While we only studied age and gender homophily here, future work may consider other forms of homophily (e.g., with respect to interests, location, or socio-economic status) or other measures of homophily, especially in a group rather than dyadic context.

Network Structure

To understand how network structure mediates trust, we calculated the following network features for each group: (1) *network density*: the number of friendships in the entire group friendship graph divided by the number of possible combinations; (2) *average clustering coefficient*: the average local clustering coefficient in the group membership graph, which measures what proportion of an individual's friends also know one another; (3) *participant degree centrality*: the number

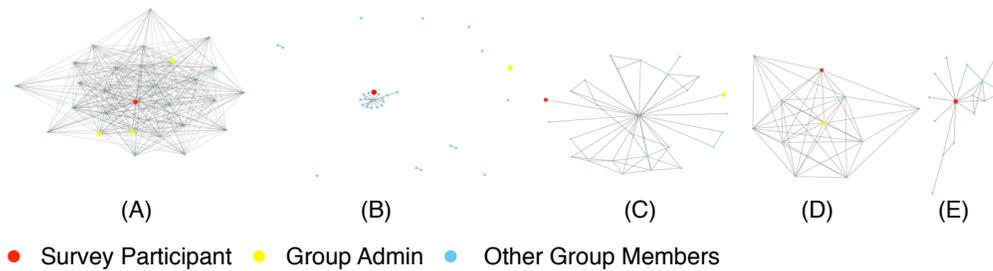


Figure 5.3: Groups differ in network density, participant degree centrality, and how a participant’s friends are linked to each other. Each node represents a group member. Each edge represents a friendship between two members. The survey participant is colored in red, and group admins are colored in yellow.

of friends a participant has in the group, normalized by group size; (4) *k-core existence*: a measure of how a participant’s friends in the group are connected with each other, calculated as whether a *k*-core component [Ugander et al., 2012] exists for participant’s friendship graph in the group (we found that $k=5$ resulted in the greatest model improvement); and (5) *average mutual friend count*: the mean number of mutual friends between participant and group members.

Figure 5.3 illustrates how several group networks in our sample differ along these network features. For example, Group A has higher network density and higher average clustering coefficient than group B. Groups C and D differ in the participant’s degree centrality. Group D contains a 5-core, but E does not.

These network features, when added to the baseline model, improves its adjusted R^2 by 0.10 ($p<.001$). Each feature was positively associated with trust in groups ($p<.001$), though we note that these network features correlate highly with one another. Considering these features separately, the average clustering coefficient was most predictive ($\beta=1.08$, $p<.001$), followed by group density ($\beta=0.93$, $p<.001$) and the participant’s degree centrality ($\beta=0.84$, $p<.001$).

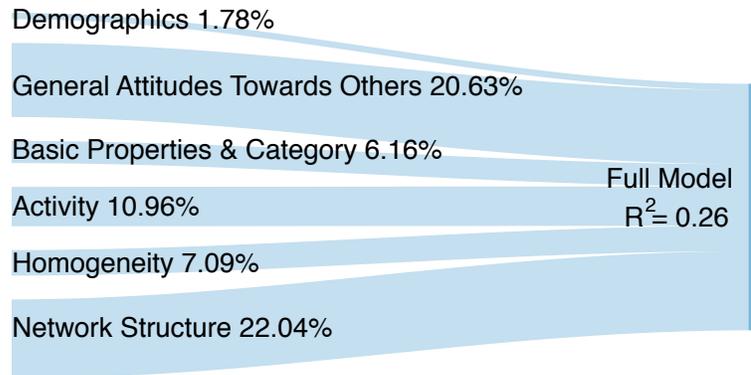


Figure 5.4: For each feature set, we calculated the average feature importance (measured by relative percent increase in MSE when a feature is removed) in predicting trust in groups. Network structure was the most important, followed by an individual’s general attitudes towards others.

5.4.3 Predicting Trust in Groups

Thus far, we have shown how various sets of group characteristics separately contribute to trust, after controlling for individual characteristics. Here, we examine how these features can together predict composite trust in groups.

A random forest model that uses all feature sets (in both Table 5.3 and Table 5.4) reached a performance of out-of-sample adjusted R^2 of 0.26 and a mean-squared error (MSE) of 0.53. We obtained similar performance using multiple linear regression.

To understand the relative importance of the different feature sets, we ranked all features by how much a random permutation of their values increased the model’s MSE. These values are shown in Figure 5.4. We find that network features are most important, followed by an individual’s general attitude towards others. Least important were demographic features. Overall, this result suggests that both individual and group characteristics are important in predicting trust in

groups.

Predicting Trust Using Only Observational Data

As we demonstrate relatively robust performance in predicting trust in groups, one might consider using such predictions to make better group or community recommendations. However, our model uses survey responses about individual differences, including disposition to trust and related concepts, to make predictions about trust in a specific group. In practical settings, it may not be feasible to administer the survey questions on individual differences to all users. This motivates the question of how well our modeling approach works in the absence of the individual differences survey features. Excluding these features, our best model obtained an adjusted R^2 of 0.15 and MSE of 0.59. In this model, network structure features were again most important, but instead followed by group activity features.

5.4.4 Group Outcomes

Theoretical accounts of trust emphasize the impact of trust on community outcomes, attributing trust to prosperity [Fukuyama, 1995], among other things. Here, we analyze how trust relates to three different group outcomes: (1) the percentage change in group size; (2) the percentage change in new tie formation (the number of new ties divided by the number of pre-existing ties) among other members of the group; and (3) the percentage change in new tie formation by the survey participant in the group. All three measures were calculated by comparing the state of the group on the day of the survey to that 28 days after. As

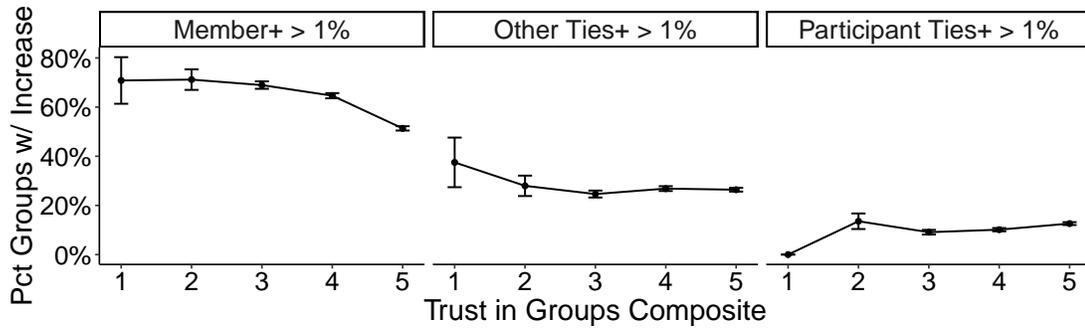


Figure 5.5: Groups with higher trust ratings are less likely to increase in size (left), more likely for the survey participant to form new connections in them (right), and had no effect on the likelihood on forming friendships among group members other than the rater (center).

these changes tend to be small, with a median change of about 1%, we instead predict whether each measure would increase by more than 1%.

Figure 5.5 shows the percentage of groups that exhibited an increase by more than 1% in each of the group outcomes listed above. Using logistic regression and controlling for basic group properties such as group size, we found that higher trust was associated with a *lower* likelihood of a group increasing in size (odds ratio -0.87, $p < .001$); and a *higher* likelihood that the survey participant would form more new friendships in the group (odds ratio 1.29, $p < .001$). Trust in a group was not predictive of the likelihood of *other* group members forming friendships in the group.

These results suggest a tension between trust and growth for online groups. Our findings are consistent with previous work on online communities that found that “cliquishness” (or high triangle density) makes a community less attractive to join and less likely to grow in size [Backstrom et al., 2006]. While membership growth is an important indicator of success for online groups [Kraut and Fiore, 2014], trust, partially elicited by small groups and exclusive member-

ship [Moser et al., 2017], can limit group expansion (but nonetheless encourages individuals to make new connections within the group). Future work can examine the relationship between trust and group longevity, as well as other interaction dynamics such as forming sub-communities within the group.

5.5 Discussion

In this work, we present a framework for predicting an individual's trust in one of their social groups on Facebook. Combining a large and diverse survey with behavioral and demographic data, we show that both individual characteristics and group characteristics contribute substantially to trust. We are able to explain a significant portion of an individual's trust in groups ($R^2=0.26$) as well as show how trust relates to outcomes such as membership growth and the formation of new within-group friendships.

This work builds on many previous studies of trust in groups by showing how features previously studied in isolation may interact with each other and how important these features are relative to each other. Beyond confirming that both an individual's general disposition to trust others [Ferguson and Peterson, 2015] and a group's size [Brewer, 1991] affect that individual's trust in a group, we further show that group size matters less to individuals with a greater disposition to trust, and that an individual's feelings of receiving social support from others in general is actually more predictive of trust in groups than their general disposition to trust. Apart from demonstrating that people do trust smaller, more private groups, we show that among groups with more than 150 members, the effect of exclusive membership decreases. Where previous

work has suggested a relation between group connectivity and trust [Coleman, 1988, Yuki et al., 2005], we also demonstrate that network measures such as the average clustering coefficient in a group are among the strongest predictors of trust in a group. Our findings on how directed communication such as likes and comments contributes to group trust corroborate similar observations in qualitative studies [Moser et al., 2017].

Nonetheless, several null results suggest areas for future exploration. While prior work suggests trust differs with sociodemographic factors [Pew, 2007], we found that age and gender explain close to zero variance in one's trust in groups. Future work may consider exploring other factors such as geography or socioeconomic status. Cultural differences may also play a significant role in trust: prior work found that an indirect relationship between two people was more likely to increase trust for Japanese than Americans [Yuki et al., 2005], suggesting that network structure may be more predictive of trust among the former. Though we found that gender- and age-homogenous groups were more trusted, we also found no evidence that gender or age homophily predicts trust in groups, in contrast to previous literature suggesting that relationships between similar individuals tend to be more trusting [Abraham et al., 2017, Ahmad et al., 2011, McPherson et al., 2001]. Understanding the extent to which these findings apply to specific situations — moms' buy-and-sell groups on Facebook are known to garner trust [Moser et al., 2017] — remains future work.

Future work may also involve investigating other potential correlates of trust such as psychological safety [Edmondson et al., 2004] and belonging [Zhao et al., 2012], as well as other outcomes of trust on online groups. For example, high trust may lead to a greater willingness to attend an event, share (or believe)

information originating from within the group, or donate to a cause.

5.5.1 Design Implications

The work reported here has several potential implications for the design of online communities.

We showed that certain types of actions (e.g., commenting and liking) are more positively associated with trust than others (e.g., posting). This adds nuance to previous findings that people have greater trust in communities in which they are more active [Cartwright and Zander, 1953]. As such, platforms could prioritize facilitating directed interactions among group members, for example, expanding features to support polling, brainstorming, and collective planning. At the same time, these findings may also inform the design of content recommendation systems. If these findings indicate that directed communication is a key signal of trust, then incorporating such signals of directed communication may better ensure that people see more content from communities that they trust more.

Consistent with prior work [Kraut and Fiore, 2014], we found that trust grows with the number of group admins and decreases with group size. As online communities grow, it may be beneficial for platforms to encourage groups to recruit additional admins to maintain existing levels of trust.

Further, network properties of online communities such as the average clustering coefficient are strong predictors of trust. Adding members that increase the average clustering of the group may be beneficial both to new members and

to the group as a whole.

Given that trust in a group correlates with behavioral signals, with additional research, platforms may also be able to provide a rough indicator of trust in groups and how it may be changing over time.

Last, our findings suggest alternative strategies for recommending groups to individuals. For instance, recommending smaller, less popular groups may not only increase the diversity of group recommendations, but also lead to greater trust and user satisfaction.

5.5.2 Limitations

Our analysis is limited to groups on Facebook. Understanding how trust differs in communities with different policies on anonymity (e.g., Reddit or Nextdoor) or that have different feedback mechanisms remains an important area for future exploration. Anonymity may increase trust by making it easier for vulnerable populations to talk about sensitive issues, but also have a disinhibiting effect and increase harassment and thus reduce trust [Kiesler et al., 1984]; indicators of reputation or popularity such as up-votes and down-votes may also influence trust, especially in the absence of other social signals [Resnick and Zeckhauser, 2002]. Still, many group properties (e.g., group size) that we examined apply to groups in general; the interactions (e.g., posting or liking) that we looked at are also common on other social media platforms. Along with the large number and diversity of groups we surveyed, we expect that many of our findings will generalize to other online communities³. While we controlled for individual

³Code to reproduce our analysis is available at <https://github.com/facebookresearch/trust-in-groups>

differences such as demographics and an individual's general attitudes towards others, understanding differences that may arise in offline groups and with regards to other factors such as location remains future work. Also, individuals may choose to join groups based on other unobserved differences (e.g., word-of-mouth). Finally, our methodology is based on correlations between variables and cannot directly suggest causation. Most significantly, it is possible that individuals who have different propensities to trust tend to join entirely different groups, explaining some of our observed differences. Similar limitations apply to the group outcomes analysis. While greater trust may lead one to connect to other members of a group, it may also arise from making these connections.

5.6 Conclusion

Groups play a significant role in an individual's social experiences and interactions. Trust, which predicts numerous positive outcomes for a group and its members, is core to a group's proper functioning. In this work, we presented a framework for predicting an individual's trust in a social group, and identified characteristics of both the individual and the group that help predict the individual's trust in the group. By surveying 6,383 Facebook Groups users about their trust attitudes and examining aggregated behavioral and demographic data for these individuals, we show that (1) an individual's propensity to trust is associated with how they trust their groups; (2) smaller, closed, older, more exclusive, or more homogeneous groups are trusted more; and (3) a group's overall friendship-network structure and an individual's position within that structure can also predict trust. Last, we demonstrate how group trust predicts outcomes at both individual and group level such as the formation of new friendship ties.

This work can contribute to future research and design decisions that better support trust in online communities and foster long-term meaningful interactions online and offline.

From a networked trust perspective, the work presented in this dissertation shows that network structure is among the most predictive factors of trust in groups. This highlights the importance of social networks in trust in networked environment. Future work can investigate how social structure contributes to trust in other platforms of digital exchange. For example, how does social structure impact interpersonal trust in the context of hiring and professional networking on LinkedIn?

At the same time, social structure can be important to understand other phenomenon related to trust. Social groups have been found to be a vehicle for misinformation [Silverman et al., 2018]. It is possible that trust in groups mediates trust in information or misinformation shared to the group. Understanding how social structure of the group impacts how people trust information shared in the group can be an important direction for future work.

CHAPTER 6

DISCUSSION AND FUTURE WORK

6.1 Algorithms of Trust and Networked Trust

This dissertation has presented three different case studies of interpersonal trust online using computational methods. In the first study, through a large-scale analysis of user-generated image cues, I built algorithms that predict image quality of products at a high accuracy (87%). A controlled experiment further showed that images selected by the algorithms outperform stock imagery in generating perceived trustworthiness in the platform. Finally, image quality predicted by the algorithm is correlated with higher sales in real world settings on eBay.

In the second study, analysis of language cues in Airbnb host profiles uncovered patterns of self-disclosure as well as how language cues lead to trust in sharing economy platforms. For example, the topic of hospitality was most effective in establishing trust, controlling for profile length. These insights show the importance of language cues and self-disclosure in establishing trust on sharing economy platforms. Further, we were able to build algorithms to predict trust in these profiles based on language cues (72% accuracy in recognizing profiles that are perceived as high or low in trustworthiness for profiles less than twenty words).

Finally, in the last study of trust, through a large-scale comprehensive study of diverse Facebook social groups, I showed that the network features are among the most predictive features of trust in social groups. People trust smaller, denser,

and more private groups. Similarly, I also developed algorithms to predict trust in social groups based on a variety individual-level and group-level features. Such algorithms can inform the ranking and recommendation for groups on Facebook to promote trust.

Taken together, these three case studies of trust have extended our understanding of interpersonal trust online in different contexts: peer-to-peer marketplaces, sharing economy platforms, and social networks. These platforms facilitate digitalized social exchange. As these platforms keep evolving, new questions about trust will continue to be raised. At the same time, there are many other platforms that facilitate social exchange and can be in scope for future investigation of trust: for example, video-based social network TikTok, Augmented Reality (AR) enabled Snapchat, etc. Trust in these platforms is likely to also play a very important role but future work is needed to understand specific contexts and risks involved.

This dissertation also argues that the term “online trust” does not sufficiently capture the factors that impact trust in networked environments. As people are embedded in social and information networks when they conduct social exchange, and as algorithms increasingly mediate such exchanges, a “networked trust” view is proposed. Details on networked trust are laid out in Section 2.3.2. As a reminder, networked trust has three focuses: (1) *cues* in Computer-Mediated Communication; (2) embeddedness in social *networks*; and (3) increasing mediation by *algorithms*. Networked trust can be a useful framework for setting future research agendas, which we discuss below.

6.2 Future Research Agenda

Here I discuss three areas of future research based on the work presented in this dissertation: (1) networked trust and misinformation; (2) AI-Mediated Communication (AI-MC); and (3) AI-Mediated Exchange Theory (AI-MET).

6.2.1 Networked Trust and Misinformation

One of the most direct applications of the networked trust framework is in understanding trust in information in misinformation research. In recent years, concerns grew over the spread of misinformation online, especially through social media [Allcott and Gentzkow, 2017, Allcott et al., 2019, Flintham et al., 2018, Grinberg et al., 2019, Flintham et al., 2018]. Although misinformation, defined as information that is false or misleading, has long existed before digital platforms, it creates new challenges now as digital platforms and algorithms optimized for engagement can result in viral diffusion of misinformation in social networks [Zhang et al., 2018]. Prior work has begun to design solutions to curb the spread of misinformation and achieved various levels of success [Zhang et al., 2018].

Using the networked trust framework, we can organize prior work on misinformation into three focuses: (1) cues: *contextual trust indicators*; (2) networks: the analysis of *how misinformation spread in networks*; and (3) algorithms: the role of *algorithms* in configuring how misinformation spreads.

The first focus, contextual trust indicators, are cues signaling the trustworthiness of information *source* [Jakesch et al., 2019, Pennycook and Rand, 2019, Zhang

et al., 2018, Kohring and Matthes, 2007]. For example, the Trust Project has designed a core set of eight Trust Indicators, including factors around the historical trustworthiness of the news outlet, the expertise of the author/reporter, and the type of work (advertisement, opinion, or news reporting)¹. The trust indicators often rely on a crowdsourcing approach, either through experts [Zhang et al., 2018] or through laypeople, which has been shown to be successful in discriminating high and low quality content and provide ratings that highly correlate with professional checkers [Pennycook and Rand, 2019, Epstein et al., 2019]. Relatedly, prior work has examined how people evaluate news credibility based on source, trust indicators, and related factors such as their own political beliefs and news literacy [Tully et al., 2019, Sundar, 1998, Jakesch et al., 2019]. On Twitter, verified status of the news distributor was not shown to have an effect on credibility of the news being shared [Vaidya et al., 2019]. Trust indicators act as cues to signal the trustworthiness of another party (usually the news source), and news consumers subsequently make decisions about trust based on these cues while considering their own propensity to trust and other factors.

The second focus, understanding how misinformation spread in networks, aligns well with the networks focus in networked trust framework. Early work on information diffusion has established that those who are exposed to a social feed in social networks are significantly more likely to spread information, and do so sooner than those who are not exposed [Bakshy et al., 2012]. In addition, weak ties play a more dominant role in the dissemination of information online, compared to strong ties [Bakshy et al., 2012]. Work on misinformation has also established patterns on how true and false information travel through networks [Vosoughi et al., 2018, Budak, 2019]. For example, false news were

¹<https://thetrustproject.org/faq/indicator>

found to diffuse “significantly farther, faster, deeper, and more broadly than the truth in all categories of information” [Vosoughi et al., 2018]. In addition, social bots were found to play a key role in the spread of low-credibility content [Shao et al., 2018]. There has also been qualitative work focused on understanding how people consume news in the context of social networks, highlighting different strategies people use to establish perceptions of trust while being embedded in networks. Although people rely on source and content to evaluate news credibility in social networks, interests in the topic also play a very important role — “when the topic of the story was not seen as personally relevant, there was little interest in figuring out whether or not it was true” [Flintham et al., 2018]. In addition, research has shown that when news is shared in social networks, trust indicators might be ignored, especially when they come from strangers rather than pre-existing social relationships [Hannak et al., 2014]. The networks focus of the above-mentioned misinformation research aligns well with the second focus of networked trust, which views interpersonal trust not in isolated dyadic interactions, but rather embedded in social and information networks where the structure of the network play an important role.

Finally, the algorithms focus investigates the role of algorithms in configuring how misinformation spread. As research on the algorithmic transparency, bias, and accountability gains wider interest [Barocas and Selbst, 2016], algorithmic transparency in the context of news media is also receiving more attention [Diakopoulos and Koliska, 2017]. Given that the majority of social media has have heavily algorithmic curated feeds [Eslami et al., 2015, Eslami et al., 2019], there is growing concern over the opacity of algorithms, especially around how such opacity affects news consumption [Diakopoulos and Koliska, 2017]. The role of algorithms in perpetuating misinformation can be found outside of social

networks as well, for example, in Amazon book recommendations² and Google search³. This algorithms focus aligns well with the third aspect of networked trust, which is concerned with understanding algorithmic mediation and how people's trust in algorithms might play an intermediating role on trust in news.

Based on this networked trust view, we can develop research questions to deepen our understanding of trust in misinformation research, especially how different focuses of networked trust interact. Instead of focusing on understanding how cues, networks, and algorithms work in isolation, we can begin to ask questions around how they interact or work together in the networked environment. For example, social contexts extracted from social networks can help improve the prediction of news credibility [Shu et al., 2019b, Shu et al., 2017, Shu et al., 2018] compared to using language cues alone [Yao et al., 2017a, Rashkin et al., 2017, Popat et al., 2016].

Below are some other examples of questions we can ask about trust in misinformation research. How does news source interact with social signals extracted from networks about the distributor (e.g., from whom the news is being shared and their social status)? How do cues (trust indicators) affect the diffusion of misinformation in social networks? Building directly off my work on trust in social groups [Ma et al., 2019a], how do people trust and interact with misinformation differently when false information is shared in groups with different network structures?

We can also ask how algorithms affect cues. How can we build personalization algorithms that are perceived as more trustworthy [Lai and Tan, 2018]?

²<https://www.wired.com/story/amazon-and-the-spread-of-health-misinformation/>

³<https://www.blog.google/around-the-globe/google-europe/fighting-disinformation-across-our-products/>

How can we build trustworthy algorithms that detect misinformation, perhaps through high transparency and explainability [Shu et al., 2019a]? These questions can help further advance our understanding of misinformation in networked environments.

6.2.2 AI-Mediated Communication (AI-MC)

The second future research direction that builds off the work presented in this dissertation is the area of AI-Mediated Communication (AI-MC). Recall that the first focus of networked trust framework is investigating how cues in Computer-Mediated Communication affect interpersonal outcomes. However, as discussed in the end of Chapter 4, algorithms can modify, augment, and even generate cues to optimize for perceived trustworthiness in online communication.

Such algorithmic ability to modify, augment, and generate cues in presentation online raises new questions about interpersonal trust and other communication outcomes. Specifically, several real-world examples have shown that AI-powered systems impact online interpersonal communication processes. For example, Google's Smart Reply, a system that generates short email responses, has already been deployed in real-world systems and accounts for 10% of the mobile replies in the Google *Inbox* [Kannan et al., 2016]. Other systems also mediate interpersonal interactions, such as profile summary auto-generation feature on LinkedIn⁴; a startup named CV Compiler that boasts using machine learning to revise resumes to boost chances of getting a high-paying job in tech⁵; and chat bots that suggests responses to other people's messages mid-conversation [Ho-

⁴<https://www.linkedin.com/pulse/solving-blank-slate-problem-through-auto-generated-summary-jalan/>

⁵<https://cvcompiler.com/>

henstein and Jung, 2018].

These AI-powered communication systems raise new questions about online interpersonal communication, especially around interpersonal trust. The term “Computer-Mediated” is no longer sufficient to capture the additional complexity that AI systems bring in interpersonal exchange. Rather than simply being a passive mediator, the AI-powered systems may alter the messages being transmitted, increasing uncertainty and risks in interpersonal communication.

To capture these new challenges brought by AI-powered communication systems, I developed the term “AI-Mediated Communication” to chart an emerging area of research. At the time of this writing, the exact definition and research questions represented by the term are still being developed [Naaman et al., 2019]. However, first attempts have already established observed effects of AI-Mediated Communication [Jakesch et al., 2019], by extending the work on computational trustworthiness of Airbnb host profiles presented in Chapter 4 [Ma et al., 2017a, Ma et al., 2017c]. Experiments show that there is a decrease in perceived trustworthiness when people are not sure whether a profile is written by human or AI [Jakesch et al., 2019]. In other words, the lack of transparency in the existence of AI hinders interpersonal trust. Future work can continue to investigate and clarify mental models around transparency in AI-Mediated Communication. Another emerging line of work in the area of AI-Mediated Communication examines how AI-suggested responses impact people’s behavior. For example, there is early evidence that positivity bias in AI-suggested responses might nudge people to respond differently to requests [Hohenstein and Jung, 2018].

Future developments in AI-Mediated Communication might also address

other important questions: For example, what characteristics of the AI systems that mediate interpersonal communication lead to acceptance and adoption? What characteristics lead to distrust and avoidance? The insights from AI-Mediated Communication will be valuable to further the emerging discussion in CHI community around design guidelines in human-AI interaction [Amer-shi et al., 2019]. Another question we can ask is how these AI systems impact decision-making. Behavior science has uncovered how people systematically made decisions that are not rational, or subject to priming and anchoring [Ariely, 2008, Kahneman, 2011]. Can AI-powered communication systems amplify or curb such patterns in decision-making? Finally, hard ethical questions need to be raised with regard to these systems, such as whether the algorithms used in AI-powered communication systems are fair, transparent, and accurate. Challenges in misrepresentation and algorithmic manipulation might exacerbate the problem of misinformation online [Susser et al., 2018]. As AI-Mediated Communication’s definition and research agenda continue to develop [Naaman et al., 2019], AI-Mediated Communication will have important implications for the design, implementation, as well as the regulation of AI-powered systems that mediate interpersonal communication online.

6.2.3 AI-Mediated Exchange Theory (AI-MET)

In the definition of AI-Mediated Communication, one of the boundary condition is that “whether a computational agent is operating on behalf of a communicator engaged in an interpersonal interaction” [Naaman et al., 2019]. Under this condition, algorithms that perform the ranking, recommendations, and classifications that support human communication are not included for consideration.

Although this boundary condition is useful for sharpening the definition and research agenda for AI-Mediated Communication, the exclusion of other types of algorithms, especially ranking and personalization algorithms, creates gaps in understanding the role of AI in digitalized exchange.

The inherent assumption of AI-Mediated Communication is that two parties are engaging in a *direct exchange* relationship through communication online. However, as I reviewed in Chapter 2, social exchange theory, the sociological theory that this dissertation builds the definition of trust on, states that there can be both *direct* and *generalized (indirect) exchanges*. Direct exchange occurs when there is a transfer of resources or information between two parties. In generalized exchange, however, often people pool resources together collectively to create greater values, and then re-distribute the values among the group [Yamagishi and Cook, 1993, Bearman, 1997]. In the digital context, open source communities and peer-to-peer platforms are considered as examples of generalized information exchange [Cheshire, 2007].

It is important, then, to consider the effect of algorithms in mediating *generalized exchange* relationships in addition to direct exchange. Future work can work on developing “AI-Mediated Exchange Theory (AI-MET)” as a framework to consider the role of AI in mediating social exchange relationships.

Specifically, the AI-Mediated Exchange Theory can discuss different mechanisms of algorithmic mediation. Defining different mechanisms through which algorithms mediate exchange can be helpful in organizing current and future work around algorithms’ impact on social interactions and society in general, especially around trust. Examples of such algorithmic mediation mechanisms might include: algorithmic aggregation, algorithmic representation, algorithmic

curation, and algorithmic augmentation.

1. *Algorithmic aggregation* can refer to the process where algorithms summarize information that is available about a potential exchange partner and make the aggregate information available to people on exchange networks. Examples of algorithmic aggregation include, reputation systems, social signals such as the number of likes, comments, and up- or down-voting.
2. *Algorithmic representation* can refer to the process where algorithms try to represent a specific entity using a label or a high dimensional vector. Examples of algorithmic representation include clustering algorithms, word embeddings, and topic modeling. Classification algorithms such as facial detection [Buolamwini and Gebru, 2018] can also be considered as examples of algorithmic representation as they represent the input data with a label (1-dimensional vector). These representations are then fed into other algorithms, such as recommendation systems, to further affect how people experience the digital platforms.
3. *Algorithmic curation* can refer to the process where algorithms select, organize and present information based on specific goals. Examples of the algorithmic curation include, news feed ranking [Eslami et al., 2015, Eslami et al., 2019], personalization and content recommendation algorithms, as well as algorithms that assist decision-making such in Applicant Tracking Systems (ATS) in hiring [Mukherjee et al., 2014, Chiang and Suen, 2015, Baldwin et al., 2014, Arya et al., 2015].
4. Finally, *algorithmic augmentation* can refer to the process where algorithms modify, augment, and even generate information related to exchange. This mechanism is essentially the focus of AI-Mediated Communication.

Examples of algorithmic augmentation include, deep fake [Bansal et al., 2018, Chan et al., 2018] for images and text, and the ability to generate realistic writings and images based on Generative Adversarial Nets [Goodfellow et al., 2014, Karras et al., 2018, Wang, 2019].

AI-Mediated Exchange Theory provides the possibility to understand how different algorithmic mediation mechanisms fit together in real-world systems to influence digitalized exchange. For example, in the hiring context, AI-Mediated Communication will be most concerned about AI augmentation — how algorithms that can improve resumes, or to generate resumes altogether might impact interpersonal outcomes in hiring. However, other algorithms such as the systems used for ranking and filtering participants can have inherent biases (e.g., gender bias [Gee, 2017]) and can affect interpersonal outcomes through AI curation. AI-MET allows for the discussion of how different mechanisms cascade to affect interpersonal outcomes. Here, we can examine how AI augmentation interacts with AI curation in the context of fairness in hiring.

In another example, in the discussion of misinformation online, AI-Mediated Communication may primarily focus on AI augmentation — understanding how people perceive others differently in the presence of AI that can generate realistic fake videos. However, as discussed above, one of the key areas of research to understand misinformation in networked environment is through understanding how misinformation spread in social networks, and the algorithms' role in recommending and amplifying misinformation in networks (AI curation). AI augmentation and curation again can cascade to affect trust in information in the context of misinformation. Again, under AI-MET, different algorithms be studied in relation to one another to understand how different algorithmic mediation

mechanisms impact interpersonal social exchange and outcomes.

In summary, three future research agenda were charted based on the work presented in this dissertation: (1) networked trust and misinformation; (2) AI-Mediated Communication (AI-MC); and (3) AI-Mediated Exchange Theory (AI-MET). These research directions either directly apply the networked trust lens in new contexts (misinformation), or extend and expand specific areas of focus of the networked trust. AI-Mediated Communication extends the first focus of networked trust, cues in Computer-Mediated Communication; and AI-Mediated Exchange Theory expands the last focus of networked trust: algorithmic mediation. These research directions will continue to address important questions about trust in the networked environment.

6.3 Conclusion

In conclusion, this dissertation presents a series of work on understanding interpersonal trust computationally, while proposing the new framework of “networked trust”. Three work presented in this dissertation detail three different computational framework on understanding and predicting trust using image cues, language cues, and social networks, in the contexts of online commerce, sharing economy, and social network groups respectively. The computational approaches help us not only gain a deeper understanding of trust in each context, but also can inform the design of digital platforms that facilitate exchange to promote trust. The algorithms of trust developed can also directly be deployed in these systems to rank and recommend exchange partners that are likely to be perceived as trustworthy to increase the overall trust on platforms.

Further, I also developed the framework of “networked trust” — a new framework for thinking about trust in hyper-connected digital environments with three focuses: (1) *cues* in Computer-Mediated Communication; (2) embeddedness in social *networks*; and (3) increasing mediation by *algorithms*. As digital platforms continue to mediate social exchange in new ways, changes to social exchange structure will continue to bring new challenges and questions about interpersonal trust in networked environments. The networked trust view can be generalized to other contexts, for example, misinformation, to help organize prior work and chart new research directions on trust in misinformation research. Other contexts where networked trust can be useful include online creative marketplaces (e.g., Spotify, Artsy, Patreon), crowdfunding platforms (e.g., Kiva, Kickstarter), and other sharing economy offerings (e.g., Airbnb Experiences). In addition, extensions of networked trust can lead to promising new areas of research, such as AI-Mediated Communication (AI-MC) and AI-Mediated Exchange Theory (AI-MET). Initial ideas for both AI-MC and AI-MET are laid out above. Future work can develop these ideas further into fuller studies and theories.

Importantly, one key limitation of this dissertation is that the flip side of trust, including distrust, bias and discrimination, has not been adequately addressed, especially in networked settings. Future work is needed to understand how digital mediation can lead to the opposite of trust, and how it leads to issues in algorithmic fairness, accountability, and transparency (FAT*). In particular, future work is needed to understand the fairness, transparency, and explainability of the algorithms predicting trust, before they are deployed in real-world systems.

Finally, to highlight how much trust has changed in the past two decades, I give one last example of social exchange that requires trust — taxi drivers.

One of the foundational book on trust, *Streetwise*, by Gambetta and Hamill and published in 2005, investigated how taxi drivers assess the trustworthiness of prospective passengers in Belfast, Northern Ireland, and New York. Through the novel use of signaling theory, they show the mechanisms taxi drivers use to establish trust rely heavily on physical cues — for example, picking up passengers only at well-lit corners and not in dark alleys [Gambetta and Hamill, 2005]. However, a little over a decade later, on Uber or Lyft, such decisions about which passengers to pick up are largely made online, through virtual networks, and increasingly controlled by algorithms [Lee et al., 2015b, Rosenblat, 2018, Rosenblat and Stark, 2016]. Real-world physical cues are discarded, while cues, networks, and algorithms become more and more important. This shift is how networked trust came into being with the digitalization of social exchange.

If anything, we are still in early stages of yet another wave of change in the digitalization of social exchange, where algorithms gain increasing importance. This dissertation has set the foundation for understanding how algorithms mediate social exchange on digital platforms. Future research can further develop AI-Mediated Exchange Theory to better understand the long-term consequences of algorithms on interpersonal outcomes in exchange, especially trust.

BIBLIOGRAPHY

- [Abrahao et al., 2017] Abrahao, B., Parigi, P., Gupta, A., and Cook, K. S. (2017). Reputation offsets trust judgments based on social biases among airbnb users. *Proceedings of the National Academy of Sciences*, 114(37):9848–9853.
- [Adamic et al., 2008] Adamic, L. A., Zhang, J., Bakshy, E., and Ackerman, M. S. (2008). Knowledge sharing and yahoo answers: everyone knows something. In *Proceedings of the International Conference on World Wide Web*, pages 665–674. ACM.
- [Ahmad et al., 2011] Ahmad, M. A., Ahmed, I., Srivastava, J., and Poole, M. S. (2011). Trust me, i’m an expert: Trust, homophily and expertise in mmos. In *Proceedings of the IEEE International Conference on Privacy, Security, Risk and Trust*, pages 882–887. IEEE.
- [Ainsworth et al., 2015] Ainsworth, M. D., Blehar, M. C., Waters, E., and Wall, S. N. (2015). *Patterns of Attachment: A Psychological Study of the Strange Situation*. Psychology Press.
- [Airbnb, 2019] Airbnb (2019). What is a superhost? <https://www.airbnb.com/help/article/828/what-is-a-superhost>. (Accessed Jul 2019).
- [Akerlof, 1978] Akerlof, G. A. (1978). The market for “lemons”: Quality uncertainty and the market mechanism. In *Uncertainty in Economics*, pages 235–251. Elsevier.
- [Allcott and Gentzkow, 2017] Allcott, H. and Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–36.
- [Allcott et al., 2019] Allcott, H., Gentzkow, M., and Yu, C. (2019). Trends in the diffusion of misinformation on social media. Technical report, National Bureau of Economic Research.
- [Amershi et al., 2019] Amershi, S., Weld, D., Vorvoreanu, M., Fournery, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., et al. (2019). Guidelines for human-ai interaction. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*.
- [Antin et al., 2011] Antin, J., Yee, R., Cheshire, C., and Nov, O. (2011). Gender

- differences in wikipedia editing. In *Proceedings of the International Symposium on Wikis and Open Collaboration*, pages 11–14. ACM.
- [Ariely, 2008] Ariely, D. (2008). *Predictably Irrational*. Harper Audio.
- [Arrow, 1974] Arrow, K. J. (1974). *The Limits of Organization*. WW Norton & Company.
- [Arya et al., 2015] Arya, D., Ha-Thuc, V., and Sinha, S. (2015). Personalized federated search at linkedin. In *Proceedings of the ACM International Conference on Information and Knowledge Management*, pages 1699–1702. ACM.
- [Ashleigh et al., 2012] Ashleigh, M. J., Higgs, M., and Dulewicz, V. (2012). A new propensity to trust scale and its relationship with individual well-being: implications for hrm policies and practices. *Human Resource Management Journal*, 22(4):360–376.
- [Bachmann, 2001] Bachmann, R. (2001). Trust, power and control in trans-organizational relations. *Organization Studies*, 22(2):337–365.
- [Backstrom et al., 2006] Backstrom, L., Huttenlocher, D., Kleinberg, J., and Lan, X. (2006). Group formation in large social networks: membership, growth, and evolution. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 44–54. ACM.
- [Backstrom and Kleinberg, 2014] Backstrom, L. and Kleinberg, J. (2014). Romantic partnerships and the dispersion of social ties: a network analysis of relationship status on facebook. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pages 831–841. ACM.
- [Bakshy et al., 2015] Bakshy, E., Messing, S., and Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on facebook. *Science*, 348(6239):1130–1132.
- [Bakshy et al., 2012] Bakshy, E., Rosenn, I., Marlow, C., and Adamic, L. (2012). The role of social networks in information diffusion. In *Proceedings of the International Conference on World Wide Web*, pages 519–528. ACM.
- [Baldwin et al., 2014] Baldwin, T., Chang, C., VanGeest, J. R., and Derezin, M. (2014). Techniques for using social proximity scores in recruiting and/or hiring. US Patent App. 14/015,751.

- [Bansal et al., 2018] Bansal, A., Ma, S., Ramanan, D., and Sheikh, Y. (2018). Recycle-gan: Unsupervised video retargeting. In *Proceedings of the European Conference on Computer Vision*, pages 119–135.
- [Barocas and Selbst, 2016] Barocas, S. and Selbst, A. D. (2016). Big data’s disparate impact. *California Law Review*, 104:671.
- [Barrera Jr and Ainlay, 1983] Barrera Jr, M. and Ainlay, S. L. (1983). The structure of social support: A conceptual and empirical analysis. *Journal of Community Psychology*, 11(2):133–143.
- [Bazarova and Choi, 2014] Bazarova, N. N. and Choi, Y. H. (2014). Self-disclosure in social media: Extending the functional approach to disclosure motivations and characteristics on social network sites. *Journal of Communication*, 64(4):635–657.
- [Bearman, 1997] Bearman, P. (1997). Generalized exchange. *American Journal of Sociology*, 102(5):1383–1415.
- [Berg et al., 1995] Berg, J., Dickhaut, J., and McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior*, 10(1):122–142.
- [Berger and Calabrese, 1975] Berger, C. R. and Calabrese, R. J. (1975). Some explorations in initial interaction and beyond: Toward a developmental theory of interpersonal communication. *Human Communication Research*, 1(2):99–112.
- [Bhattacharya et al., 2010] Bhattacharya, S., Sukthankar, R., and Shah, M. (2010). A framework for photo-quality assessment and enhancement based on visual aesthetics. In *Proceedings of the ACM International Conference on Multimedia*, pages 271–280. ACM.
- [Bijlsma and Koopman, 2003] Bijlsma, K. and Koopman, P. (2003). Introduction: trust within organisations. *Personnel Review*, 32(5):543–555.
- [Bjørnskov, 2007] Bjørnskov, C. (2007). Determinants of generalized trust: A cross-country comparison. *Public Choice*, 130(1-2):1–21.
- [Bland et al., 2007] Bland, E. M., Black, G. S., and Lawrimore, K. (2007). Risk-reducing and risk-enhancing factors impacting online auction outcomes: empirical evidence from ebay auctions. *Electronic Commerce Research*, 8(4):236.
- [Blau and Kahn, 2007] Blau, F. D. and Kahn, L. M. (2007). The gender pay gap:

- Have women gone as far as they can? *Academy of Management Perspectives*, 21(1):7–23.
- [Blau, 1964] Blau, P. (1964). *Exchange and Power in Social Life*. John Wiley & Sons.
- [Bolukbasi et al., 2016] Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, pages 4349–4357.
- [Boss, 1978] Boss, R. W. (1978). Trust and managerial problem solving revisited. *Group & Organization Studies*, 3(3):331–342.
- [Bowlby, 1969] Bowlby, J. (1969). *Attachment and Loss: Attachment*. Basic books.
- [boyd and Ellison, 2007] boyd, d. and Ellison, N. (2007). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 13(1):210–230.
- [Boyd and Ellison, 2007] Boyd, D. M. and Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 13(1):210–230.
- [Bradley and Terry, 1952] Bradley, R. A. and Terry, M. E. (1952). Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.
- [Brewer, 1991] Brewer, M. B. (1991). The social self: On being the same and different at the same time. *Personality and Social Psychology Bulletin*, 17(5):475–482.
- [Budak, 2019] Budak, C. (2019). What happened? the spread of fake news publisher content during the 2016 us presidential election. In *Proceedings of the International Conference on World Wide Web*, pages 139–150. ACM.
- [Buolamwini and Gebru, 2018] Buolamwini, J. and Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 77–91.
- [Butler Jr, 1999] Butler Jr, J. K. (1999). Trust expectations, information sharing,

- climate of trust, and negotiation effectiveness and efficiency. *Group & Organization Management*, 24(2):217–238.
- [Cartwright and Zander, 1953] Cartwright, D. and Zander, A. (1953). Group cohesiveness: introduction. *Group Dynamics: Research and Theory*. Evanston, IL: Row Peterson.
- [Chamberlain, 2016] Chamberlain, A. (2016). Demystifying the gender pay gap. *Mill Valley, CA: Glassdoor*.
- [Chan et al., 2018] Chan, C., Ginosar, S., Zhou, T., and Efros, A. A. (2018). Everybody dance now. *arXiv preprint arXiv:1808.07371*.
- [Chanley et al., 2000] Chanley, V. A., Rudolph, T. J., and Rahn, W. M. (2000). The origins and consequences of public trust in government: A time series analysis. *Public Opinion Quarterly*, 64(3):239–256.
- [Chaudhuri and Holbrook, 2001] Chaudhuri, A. and Holbrook, M. B. (2001). The chain of effects from brand trust and brand affect to brand performance: the role of brand loyalty. *Journal of Marketing*, 65(2):81–93.
- [Cheng et al., 2014] Cheng, J., Adamic, L., Dow, P. A., Kleinberg, J. M., and Leskovec, J. (2014). Can cascades be predicted? In *Proceedings of the International Conference on World Wide Web*, pages 925–936. ACM.
- [Cheshire, 2007] Cheshire, C. (2007). Selective incentives and generalized information exchange. *Social Psychology Quarterly*, 70(1):82–100.
- [Cheshire, 2011] Cheshire, C. (2011). Online trust, trustworthiness, or assurance? *Daedalus*, 140(4):49–58.
- [Cheshire and Cook, 2004] Cheshire, C. and Cook, K. S. (2004). The emergence of trust networks under uncertainty-implications for internet interactions. *Analyse und Kritik*, 26(1):220.
- [Chiang and Suen, 2015] Chiang, J. K.-H. and Suen, H.-Y. (2015). Self-presentation and hiring recommendations in online communities: Lessons from linkedin. *Computers in Human Behavior*, 48:516–524.
- [Choi and Mai, 2018] Choi, Y. and Mai, D. Q. (2018). The sustainable role of the e-trust in the b2c e-commerce of vietnam. *Sustainability*, 10(1).

- [Chung et al., 2012] Chung, S. H., Goswami, A., Lee, H., and Hu, J. (2012). The impact of images on user clicks in product search. In *Proceedings of the International Workshop on Multimedia Data Mining*, pages 25–33. ACM.
- [Coleman, 1988] Coleman, J. S. (1988). Social capital in the creation of human capital. *American Journal of Sociology*, 94:S95–S120.
- [Colquitt et al., 2011] Colquitt, J. A., LePine, J. A., Zapata, C. P., and Wild, R. E. (2011). Trust in typical and high-reliability contexts: Building and reacting to trust among firefighters. *Academy of Management Journal*, 54(5):999–1015.
- [Colquitt et al., 2007] Colquitt, J. A., Scott, B. A., and LePine, J. A. (2007). Trust, trustworthiness, and trust propensity: a meta-analytic test of their unique relationships with risk taking and job performance. *Journal of Applied Psychology*, 92(4):909.
- [Cook et al., 2009] Cook, K. S., Snijders, C., Buskens, V., and Cheshire, C. (2009). *eTrust: Forming Relationships in the Online World: Forming Relationships in the Online World*. Russell Sage Foundation.
- [Cook and Whitmeyer, 1992] Cook, K. S. and Whitmeyer, J. M. (1992). Two approaches to social structure: Exchange theory and network analysis. *Annual Review of Sociology*, 18(1):109–127.
- [Cook et al., 2005] Cook, K. S., Yamagishi, T., Cheshire, C., Cooper, R., Matsuda, M., and Mashima, R. (2005). Trust building via risk taking: A cross-societal experiment. *Social Psychology Quarterly*, 68(2):121–142.
- [Corbitt et al., 2003] Corbitt, B. J., Thanasankit, T., and Yi, H. (2003). Trust and e-commerce: a study of consumer perceptions. *Electronic Commerce Research and Applications*, 2(3):203–215.
- [Corritore et al., 2001] Corritore, C. L., Wiedenbeck, S., and Kracher, B. (2001). The elements of online trust. In *Companion to the ACM Conference on Human Factors in Computing Systems*, pages 504–505. ACM.
- [Costa et al., 2001] Costa, A. C., Roe, R. A., and Taillieu, T. (2001). Trust within teams: The relation with performance effectiveness. *European Journal of Work and Organizational Psychology*, 10(3):225–244.
- [Cozby, 1972] Cozby, P. C. (1972). Self-disclosure, reciprocity and liking. *Sociometry*, pages 151–160.

- [Danescu-Niculescu-Mizil et al., 2013a] Danescu-Niculescu-Mizil, C., Sudhof, M., Jurafsky, D., Leskovec, J., and Potts, C. (2013a). A computational approach to politeness with application to social factors. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.
- [Danescu-Niculescu-Mizil et al., 2013b] Danescu-Niculescu-Mizil, C., Sudhof, M., Jurafsky, D., Leskovec, J., and Potts, C. (2013b). A computational approach to politeness with application to social factors. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.
- [Datta et al., 2018] Datta, A., Makagon, J., Mulligan, D., and Tschantz, M. (2018). Discrimination in online advertising a multidisciplinary inquiry. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*.
- [Datta et al., 2006] Datta, R., Joshi, D., Li, J., and Wang, J. Z. (2006). Studying aesthetics in photographic images using a computational approach. In *Proceedings of the European Conference on Computer Vision*, pages 288–301. Springer.
- [Delgado-Ballester and Luis Munuera-Alemán, 2001] Delgado-Ballester, E. and Luis Munuera-Alemán, J. (2001). Brand trust in the context of consumer loyalty. *European Journal of Marketing*, 35(11/12):1238–1258.
- [Deng et al., 2009] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE.
- [Denson et al., 2006] Denson, T. F., Lickel, B., Curtis, M., Stenstrom, D. M., and Ames, D. R. (2006). The roles of entitativity and essentiality in judgments of collective responsibility. *Group Processes & Intergroup Relations*, 9(1):43–61.
- [Denters, 2002] Denters, B. (2002). Size and political trust: evidence from denmark, the netherlands, norway, and the united kingdom. *Environment and Planning C: Government and Policy*, 20(6):793–812.
- [Derlega et al., 1987] Derlega, V. J., Winstead, B. A., Wong, P., and Greenspan, M. (1987). Self-disclosure and relationship development: An attributional analysis. *Interpersonal Processes: New Directions in Communication Research*, pages 172–187.
- [DeSteno et al., 2012] DeSteno, D., Breazeal, C., Frank, R. H., Pizarro, D., Baumann, J., Dickens, L., and Lee, J. J. (2012). Detecting the trustworthiness of novel partners in economic exchange. *Psychological Science*, 23(12):1549–1556.

- [Di et al., 2014] Di, W., Sundaresan, N., PIRAMUTHU, R., and Bhardwaj, A. (2014). Is a picture really worth a thousand words? on the role of images in e-commerce. In *Proceedings of the ACM International Conference on Web Search and Data Mining*, pages 633–642. ACM.
- [Diakopoulos and Koliska, 2017] Diakopoulos, N. and Koliska, M. (2017). Algorithmic transparency in the news media. *Digital Journalism*, 5(7):809–828.
- [Dirks, 1999] Dirks, K. T. (1999). The effects of interpersonal trust on work group performance. *Journal of Applied Psychology*, 84(3):445.
- [Dirks and Ferrin, 2002] Dirks, K. T. and Ferrin, D. L. (2002). Trust in leadership: Meta-analytic findings and implications for research and practice. *Journal of Applied Psychology*, 87(4):611.
- [Doleac and Stein, 2013] Doleac, J. L. and Stein, L. C. (2013). The visible hand: Race and online market outcomes. *The Economic Journal*, 123(572).
- [Donath, 2007] Donath, J. (2007). Signals in social supernets. *Journal of Computer-Mediated Communication*, 13(1):231–251.
- [Dubey et al., 2017] Dubey, A., Abhinav, K., Hamilton, M., and Kass, A. (2017). Analyzing gender pay gap in freelancing marketplace. In *Proceedings of the ACM Conference on Computers and People Research*, pages 13–19. ACM.
- [Dunbar, 1992] Dunbar, R. I. (1992). Neocortex size as a constraint on group size in primates. *Journal of Human Evolution*, 22(6):469–493.
- [Dwyer et al., 2007] Dwyer, C., Hiltz, S., and Passerini, K. (2007). Trust and privacy concern within social networking sites: A comparison of facebook and myspace. *Proceedings of the Americas Conference on Information Systems*, page 339.
- [Easley et al., 2010] Easley, D., Kleinberg, J., et al. (2010). *Networks, Crowds, and Markets*, volume 8. Cambridge University Press.
- [eBay, 2019] eBay (2019). Photo tips. <https://pages.ebay.com/seller-center/listing-and-marketing/photo-tips.html>. (Accessed Jul 2019).
- [Eckhouse et al., 2019] Eckhouse, L., Lum, K., Conti-Cook, C., and Ciccolini, J.

- (2019). Layers of bias: A unified approach for understanding problems with risk assessment. *Criminal Justice and Behavior*, 46(2):185–209.
- [Edelman et al., 2017] Edelman, B., Luca, M., and Svirsky, D. (2017). Racial discrimination in the sharing economy: Evidence from a field experiment. *American Economic Journal: Applied Economics*, 9(2):1–22.
- [Edelman and Luca, 2014] Edelman, B. G. and Luca, M. (2014). Digital discrimination: The case of airbnb. com. *Harvard Business School NOM Unit Working Paper*, (14-054).
- [Edmondson et al., 2004] Edmondson, A. C., Kramer, R. M., and Cook, K. S. (2004). Psychological safety, trust, and learning in organizations: A group-level lens. *Trust and Distrust in Organizations: Dilemmas and Approaches*, 12:239–272.
- [Ellison and Hancock, 2013] Ellison, N. B. and Hancock, J. T. (2013). Profile as promise: Honest and deceptive signals in online dating. *IEEE Security and Privacy*, 11(5):84–88.
- [Ellison et al., 2012] Ellison, N. B., Hancock, J. T., and Toma, C. L. (2012). Profile as promise: A framework for conceptualizing veracity in online dating self-presentations. *New Media & Society*, 14(1):45–62.
- [Ellison et al., 2014] Ellison, N. B., Vitak, J., Gray, R., and Lampe, C. (2014). Cultivating social resources on social network sites: Facebook relationship maintenance behaviors and their role in social capital processes. *Journal of Computer-Mediated Communication*, 19(4):855–870.
- [Ely, 1980] Ely, J. H. (1980). *Democracy and Distrust: A Theory of Judicial Review*. Harvard University Press.
- [Emerson, 1976] Emerson, R. M. (1976). Social exchange theory. *Annual Review of Sociology*, pages 335–362.
- [Epstein et al., 2019] Epstein, Z., Pennycook, G., and Rand, D. (2019). Letting the crowd steer the algorithm: Laypeople can effectively identify misinformation sources. PsyArXiv. Working Paper.
- [Ert et al., 2016] Ert, E., Fleischer, A., and Magen, N. (2016). Trust and reputation in the sharing economy: The role of personal photos in airbnb. *Tourism Management*, 55:62–73.

- [Eslami et al., 2015] Eslami, M., Rickman, A., Vaccaro, K., Aleyasen, A., Vuong, A., Karahalios, K., Hamilton, K., and Sandvig, C. (2015). I always assumed that i wasn't really that close to [her]: Reasoning about invisible algorithms in news feeds. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pages 153–162. ACM.
- [Eslami et al., 2019] Eslami, M., Vaccaro, K., Gilbert, E., Lee, M. K., and Karahalios, K. (2019). User attitudes towards algorithmic opacity and transparency in online reviewing platforms. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*. ACM.
- [Facebook, 2018] Facebook (2018). Facebook help center. <https://www.facebook.com/help/1629740080681586>. (Accessed Sep 2018).
- [Ferguson and Peterson, 2015] Ferguson, A. J. and Peterson, R. S. (2015). Sinking slowly: Diversity in propensity to trust predicts downward trust spirals in small groups. *Journal of Applied Psychology*, 100(4):1012.
- [Filippas et al., 2018] Filippas, A., Horton, J., and Golden, J. (2018). Reputation inflation: Evidence from an online labor market. In *Proceedings of the ACM Conference on Economics and Computation*.
- [Fine and Holyfield, 1996] Fine, G. A. and Holyfield, L. (1996). Secrecy, trust, and dangerous leisure: Generating group cohesion in voluntary organizations. *Social Psychology Quarterly*, pages 22–38.
- [Flintham et al., 2018] Flintham, M., Karner, C., Bachour, K., Creswick, H., Gupta, N., and Moran, S. (2018). Falling for fake news: investigating the consumption of news via social media. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, page 376. ACM.
- [Fogel and Nehmad, 2009] Fogel, J. and Nehmad, E. (2009). Internet social network communities: Risk taking, trust, and privacy concerns. *Computers in Human Behavior*, 25(1):153–160.
- [Forest and Wood, 2012] Forest, A. L. and Wood, J. V. (2012). When social networking is not working: Individuals with low self-esteem recognize but do not reap the benefits of self-disclosure on facebook. *Psychological Science*, 23(3):295–302.
- [Fradkin et al., 2015] Fradkin, A., Grewal, E., Holtz, D., and Pearson, M. (2015). Bias and reciprocity in online reviews: Evidence from field experiments on

- airbnb. In *Proceedings of the ACM Conference on Economics and Computation*, pages 641–641. ACM.
- [Fukuyama, 1995] Fukuyama, F. (1995). *Trust: The Social Virtues and the Creation of Prosperity*. Free Press.
- [Gambetta, 1988] Gambetta, D. (1988). *Trust: Making and Breaking Cooperative Relations*. Basil Blackwell.
- [Gambetta, 2009] Gambetta, D. (2009). *Codes of the Underworld: How Criminals Communicate*. Princeton University Press.
- [Gambetta and Hamill, 2005] Gambetta, D. and Hamill, H. (2005). *Streetwise: How Taxi Drivers Establish Customer’s Trustworthiness*. Russell Sage Foundation.
- [Gee, 2017] Gee, K. (2017). In unilever’s radical hiring experiment, resumes are out, algorithms are in. *Wall Street Journal*, 26.
- [Gefen and Straub, 2004] Gefen, D. and Straub, D. W. (2004). Consumer trust in b2c e-commerce and the importance of social presence: experiments in e-products and e-services. *Omega*, 32(6):407–424.
- [Gibbs et al., 2010] Gibbs, J. L., Ellison, N. B., and Lai, C.-H. (2010). First comes love, then comes google: An investigation of uncertainty reduction strategies and self-disclosure in online dating. *Communication Research*, page 0093650210377091.
- [Gilbert and Karahalios, 2009] Gilbert, E. and Karahalios, K. (2009). Predicting tie strength with social media. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pages 211–220. ACM.
- [Gill et al., 2005] Gill, H., Boies, K., Finegan, J. E., and McNally, J. (2005). Antecedents of trust: Establishing a boundary condition for the relation between propensity to trust and intention to trust. *Journal of Business and Psychology*, 19(3):287–302.
- [Glaeser et al., 2000] Glaeser, E. L., Laibson, D. I., Scheinkman, J. A., and Soutter, C. L. (2000). Measuring trust. *The Quarterly Journal of Economics*, 115(3):811–846.
- [Goodfellow et al., 2016] Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. <http://www.deeplearningbook.org>.

- [Goodfellow et al., 2014] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680.
- [Google, 2016] Google (2016). Google translate api documentation. <https://cloud.google.com/translate/>.
- [Google, 2019] Google (2019). Google merchant center: Image guidelines. https://support.google.com/merchants/answer/6324350?hl=en&ref_topic=6324338. (Accessed Jul 2019).
- [Goswami et al., 2011] Goswami, A., Chittar, N., and Sung, C. H. (2011). A study on the impact of product images on user clicks for online shopping. In *Companion to the International Conference on World Wide Web*, pages 45–46. ACM.
- [Grabner-Kraeuter, 2002] Grabner-Kraeuter, S. (2002). The role of consumers' trust in online-shopping. *Journal of Business Ethics*, 39(1-2):43–50.
- [Granovetter, 1985] Granovetter, M. (1985). Economic action and social structure: The problem of embeddedness. *American Journal of Sociology*, 91(3):481–510.
- [Granovetter, 1973] Granovetter, M. S. (1973). The strength of weak ties. *American Journal of Sociology*, 78(6):1360–1380.
- [Grbovic and Cheng, 2018] Grbovic, M. and Cheng, H. (2018). Real-time personalization using embeddings for search ranking at airbnb. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 311–320. ACM.
- [Grinberg et al., 2019] Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., and Lazer, D. (2019). Fake news on twitter during the 2016 us presidential election. *Science*, 363(6425):374–378.
- [Gubin et al., 2017] Gubin, M., Kao, W., Vickrey, D., and Maykov, A. (2017). News feed ranking model based on social information of viewer. US Patent 9,582,786.
- [Guillory and Hancock, 2012] Guillory, J. and Hancock, J. T. (2012). The effect of linkedin on deception in resumes. *Cyberpsychology, Behavior, and Social Networking*, 15(3):135–140.

- [Gulati, 1995] Gulati, R. (1995). Does familiarity breed trust? the implications of repeated ties for contractual choice in alliances. *Academy of Management Journal*, 38(1):85–112.
- [Gunther, 1988] Gunther, A. (1988). Attitude extremity and trust in media. *Journalism Quarterly*, 65(2):279–287.
- [Guttentag, 2015] Guttentag, D. (2015). Airbnb: Disruptive innovation and the rise of an informal tourism accommodation sector. *Current Issues in Tourism*, 18(12):1192–1217.
- [Hamari et al., 2016] Hamari, J., Sjöklint, M., and Ukkonen, A. (2016). The sharing economy: Why people participate in collaborative consumption. *Journal of the Association for Information Science and Technology*, 67(9):2047–2059.
- [Hannak et al., 2014] Hannak, A., Margolin, D., Keegan, B., and Weber, I. (2014). Get back! you don't know me like that: The social mediation of fact checking interventions in twitter conversations. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*.
- [Hannak et al., 2017] Hannak, A., Wagner, C., Garcia, D., Mislove, A., Strohmaier, M., and Wilson, C. (2017). Bias in online freelance marketplaces: Evidence from taskrabbit and fiverr. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pages 1914–1933. ACM.
- [Hardin, 1999] Hardin, R. (1999). Do we want trust in government. *Democracy and Trust*, pages 22–41.
- [Hardin, 2002] Hardin, R. (2002). *Trust and Trustworthiness*. Russell Sage Foundation.
- [Hether et al., 2014] Hether, H. J., Murphy, S. T., and Valente, T. W. (2014). It's better to give than to receive: The role of social support, trust, and participation on health-related social networking sites. *Journal of Health Communication*, 19(12):1424–1439.
- [Hogg, 1993] Hogg, M. A. (1993). Group cohesiveness: A critical review and some new directions. *European Review of Social Psychology*, 4(1):85–111.
- [Hohenstein and Jung, 2018] Hohenstein, J. and Jung, M. (2018). Ai-supported messaging: An investigation of human-human text conversation with ai sup-

- port. In *Companion to the ACM Conference on Human Factors in Computing Systems*, page LBW089. ACM.
- [Holtz et al., 2017] Holtz, D., Lynn MacLean, D., and Aral, S. (2017). Social structure and trust in massive digital markets.
- [Homans, 1958] Homans, G. C. (1958). Social behavior as exchange. *American Journal of Sociology*, pages 597–606.
- [Hong and Cho, 2011] Hong, I. B. and Cho, H. (2011). The impact of consumer trust on attitudinal loyalty and purchase intentions in b2c e-marketplaces: Intermediary trust vs. seller trust. *International Journal of Information Management*, 31(5):469–479.
- [Hutson et al., 2018] Hutson, J. A., Taft, J. G., Barocas, S., and Levy, K. (2018). Debiasing desire: Addressing bias & discrimination on intimate platforms. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):73.
- [Ikkala and Lampinen, 2015] Ikkala, T. and Lampinen, A. (2015). Monetizing network hospitality: Hospitality and sociability in the context of airbnb. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pages 1033–1044. ACM.
- [Inside Airbnb., 2016] Inside Airbnb. (2016). About inside airbnb. <http://insideairbnb.com/about.html>. (Accessed May 2016).
- [Jakesch et al., 2019] Jakesch, M., French, M., Ma, X., Hancock, J. T., and Naaman, M. (2019). Ai-mediated communication: How the perception that profile text was written by ai affects trustworthiness. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, volume 51, page 22.
- [Jarvenpaa and Leidner, 1999] Jarvenpaa, S. L. and Leidner, D. E. (1999). Communication and trust in global virtual teams. *Organization Science*, 10(6):791–815.
- [Johnson and Mislin, 2011] Johnson, N. D. and Mislin, A. A. (2011). Trust games: A meta-analysis. *Journal of Economic Psychology*, 32(5):865–889.
- [Johnson-George and Swap, 1982] Johnson-George, C. and Swap, W. C. (1982). Measurement of specific interpersonal trust: Construction and validation of a scale to assess trust in a specific other. *Journal of Personality and Social Psychology*, 43(6):1306.

- [Jones and Leonard, 2008] Jones, K. and Leonard, L. N. (2008). Trust in consumer-to-consumer electronic commerce. *Information & Management*, 45(2):88–95.
- [Jøsang et al., 2007] Jøsang, A., Ismail, R., and Boyd, C. (2007). A survey of trust and reputation systems for online service provision. *Decision Support Systems*, 43(2):618–644.
- [Jourard and Lasakow, 1958] Jourard, S. M. and Lasakow, P. (1958). Some factors in self-disclosure. *The Journal of Abnormal and Social Psychology*, 56(1):91.
- [Kahneman, 2011] Kahneman, D. (2011). *Thinking, Fast and Slow*. Macmillan.
- [Kannan et al., 2016] Kannan, A., Kurach, K., Ravi, S., Kaufmann, T., Tomkins, A., Miklos, B., Corrado, G., Lukacs, L., Ganea, M., Young, P., et al. (2016). Smart reply: Automated response suggestion for email. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 955–964. ACM.
- [Kantsperger and Kunz, 2010] Kantsperger, R. and Kunz, W. H. (2010). Consumer trust in service companies: a multiple mediating analysis. *Managing Service Quality: An International Journal*, 20(1):4–25.
- [Karras et al., 2018] Karras, T., Laine, S., and Aila, T. (2018). A style-based generator architecture for generative adversarial networks. *arXiv preprint arXiv:1812.04948*.
- [Kiapour et al., 2015] Kiapour, M. H., Han, X., Lazebnik, S., Berg, A. C., and Berg, T. L. (2015). Where to buy it: Matching street clothing photos in online shops. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*. IEEE.
- [Kiesler et al., 1984] Kiesler, S., Siegel, J., and McGuire, T. W. (1984). Social psychological aspects of computer-mediated communication. *American Psychologist*, 39(10):1123.
- [Kioussis, 2001] Kioussis, S. (2001). Public trust or mistrust? perceptions of media credibility in the information age. *Mass Communication & Society*, 4(4):381–403.
- [Kiyonari et al., 2006] Kiyonari, T., Yamagishi, T., Cook, K. S., and Cheshire, C. (2006). Does trust beget trustworthiness? trust and trustworthiness in

- two games and two cultures: A research note. *Social Psychology Quarterly*, 69(3):270–283.
- [Kizilcec, 2016] Kizilcec, R. F. (2016). How much information?: Effects of transparency on trust in an algorithmic interface. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pages 2390–2395. ACM.
- [Kohring and Matthes, 2007] Kohring, M. and Matthes, J. (2007). Trust in news media: Development and validation of a multidimensional scale. *Communication Research*, 34(2):231–252.
- [Kramer et al., 2014] Kramer, A. D., Guillory, J. E., and Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. volume 111, pages 8788–8790. National Academy of Sciences.
- [Kramer, 1999] Kramer, R. M. (1999). Trust and distrust in organizations: Emerging perspectives, enduring questions. *Annual Review of Psychology*, 50(1):569–598.
- [Kraut and Fiore, 2014] Kraut, R. E. and Fiore, A. T. (2014). The role of founders in building online groups. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pages 722–732. ACM.
- [Kraut and Resnick, 2012] Kraut, R. E. and Resnick, P. (2012). *Building Successful Online Communities: Evidence-Based Social Design*. MIT Press.
- [Kricheli-Katz and Regev, 2016] Kricheli-Katz, T. and Regev, T. (2016). How many cents on the dollar? women and men in product markets. *Science Advances*, 2(2):e1500599.
- [Kuwabara, 2015] Kuwabara, K. (2015). Do reputation systems undermine trust? divergent effects of enforcement type on generalized trust and trustworthiness¹. *American Journal of Sociology*, 120(5):1390–1428.
- [La Macchia et al., 2016] La Macchia, S. T., Louis, W. R., Hornsey, M. J., and Leonardelli, G. J. (2016). In small we trust: Lay theories about small and large groups. *Personality and Social Psychology Bulletin*, 42(10):1321–1334.
- [Lai and Tan, 2018] Lai, V. and Tan, C. (2018). On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*.

- [Lampe et al., 2006] Lampe, C., Ellison, N., and Steinfield, C. (2006). A face (book) in the crowd: Social searching vs. social browsing. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pages 167–170. ACM.
- [Lampinen and Cheshire, 2016] Lampinen, A. and Cheshire, C. (2016). Hosting via airbnb: Motivations and financial assurances in monetized network hospitality. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pages 1669–1680. ACM.
- [Lankton and McKnight, 2011] Lankton, N. K. and McKnight, D. H. (2011). What does it mean to trust facebook?: examining technology and interpersonal trust beliefs. *ACM SIGMIS Database: The DATABASE for Advances in Information Systems*, 42(2):32–54.
- [Larzelere and Huston, 1980] Larzelere, R. E. and Huston, T. L. (1980). The dyadic trust scale: Toward understanding interpersonal trust in close relationships. *Journal of Marriage and the Family*, pages 595–604.
- [Lavergne and Mullainathan, 2004] Lavergne, M. and Mullainathan, S. (2004). Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. *The American Economic Review*, 94(4):991–1013.
- [Lee et al., 2015a] Lee, D., Hyun, W., Ryu, J., Lee, W. J., Rhee, W., and Suh, B. (2015a). An analysis of social features associated with room sales of airbnb. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work & Social Computing*. ACM.
- [Lee et al., 2013] Lee, J. J., Knox, B., Baumann, J., Breazeal, C., and DeSteno, D. (2013). Computationally modeling interpersonal trust. *Frontiers in Psychology*, 4:893.
- [Lee et al., 2015b] Lee, M. K., Kusbit, D., Metsky, E., and Dabbish, L. (2015b). Working with machines: The impact of algorithmic and data-driven management on human workers. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pages 1603–1612. ACM.
- [Levy and Barocas, 2017] Levy, K. and Barocas, S. (2017). Designing against discrimination in online markets. *Berkeley Technology Law Journal*.
- [Li et al., 2010] Li, C., Loui, A. C., and Chen, T. (2010). Towards aesthetics: A

- photo quality assessment and photo selection system. In *Proceedings of the ACM International Conference on Multimedia*, pages 827–830. ACM.
- [Linden et al., 2003] Linden, G., Smith, B., and York, J. (2003). Amazon. com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing*, (1):76–80.
- [Liu et al., 2017] Liu, M., Ding, W., Park, D. H., Fang, Y., Yan, R., and Hu, X. (2017). Which used product is more sellable? a time-aware approach. *Information Retrieval Journal*, 20(2):81–108.
- [Liu et al., 2016] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., and Berg, A. C. (2016). Ssd: Single shot multibox detector. In *Proceedings of the European Conference on Computer Vision*, pages 21–37. Springer.
- [Lu et al., 2014] Lu, X., Lin, Z., Jin, H., Yang, J., and Wang, J. Z. (2014). Rapid: Rating pictorial aesthetics using deep learning. In *Proceedings of the ACM International Conference on Multimedia*, pages 457–466. ACM.
- [Luhmann, 1979] Luhmann, N. (1979). *Trust and Power: Two Works by Niklas Luhmann*. Chichester: John Wiley.
- [Ma et al., 2019a] Ma, X., Cheng, J., Iyer, S., and Naaman, M. (2019a). When do people trust their social groups? In *Proceedings of the ACM Conference on Human Factors in Computing Systems*. ACM.
- [Ma et al., 2016] Ma, X., Hancock, J., and Naaman, M. (2016). Anonymity, intimacy and self-disclosure in social media. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pages 3857–3869. ACM.
- [Ma et al., 2017a] Ma, X., Hancock, J. T., Lim Mingjie, K., and Naaman, M. (2017a). Self-disclosure and perceived trustworthiness of airbnb host profiles. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work & Social Computing*. ACM.
- [Ma et al., 2017b] Ma, X., Hancock, J. T., Mingjie, K. L., and Naaman, M. (2017b). Self-disclosure and perceived trustworthiness of airbnb host profiles. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work & Social Computing*. ACM.
- [Ma et al., 2019b] Ma, X., Mezghani, L., Wilber, K., Hong, H., Piramuthu, R., Naaman, M., and Belongie, S. (2019b). Understanding image quality and trust

- in peer-to-peer marketplaces. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, pages 511–520. IEEE.
- [Ma et al., 2017c] Ma, X., Neeraj, T., and Naaman, M. (2017c). A computational approach to perceived trustworthiness of airbnb host profiles. In *Eleventh International AAAI Conference on Web and Social Media*. AAAI.
- [Marsh and Dibben, 2005] Marsh, S. and Dibben, M. R. (2005). Trust, untrust, distrust and mistrust—an exploration of the dark (er) side. In *Proceedings of the International Conference on Trust Management*, pages 17–33. Springer.
- [Marwick and Boyd, 2011] Marwick, A. E. and Boyd, D. (2011). I tweet honestly, i tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society*, 13(1):114–133.
- [Mayer and Davis, 1999] Mayer, R. C. and Davis, J. H. (1999). The effect of the performance appraisal system on trust for management: A field quasi-experiment. *Journal of Applied Psychology*, 84(1):123.
- [Mayer et al., 1995] Mayer, R. C., Davis, J. H., and Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review*, 20(3):709–734.
- [Mayer and Gavin, 2005] Mayer, R. C. and Gavin, M. B. (2005). Trust in management and performance: who minds the shop while the employees watch the boss? *Academy of Management Journal*, 48(5):874–888.
- [McEvily et al., 2003] McEvily, B., Perrone, V., and Zaheer, A. (2003). Trust as an organizing principle. *Organization Science*, 14(1):91–103.
- [McPherson et al., 2001] McPherson, M., Smith-Lovin, L., and Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1):415–444.
- [Merrill and Cheshire, 2017] Merrill, N. and Cheshire, C. (2017). Trust your heart: Assessing cooperation and trust with biosignals in computer-mediated interactions. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pages 2–12. ACM.
- [Meyerson et al., 1996] Meyerson, D., Weick, K. E., and Kramer, R. M. (1996). Swift trust and temporary groups. *Trust in Organizations: Frontiers of Theory and Research*, 166:195.

- [Michailidou et al., 2008] Michailidou, E., Harper, S., and Bechhofer, S. (2008). Visual complexity and aesthetic perception of web pages. In *Proceedings of the ACM International Conference on Design of Communication*, pages 215–224. ACM.
- [Miller, 1974] Miller, A. H. (1974). Political issues and trust in government: 1964–1970. *American Political Science Review*, 68(03):951–972.
- [Miller and Mitamura, 2003] Miller, A. S. and Mitamura, T. (2003). Are surveys on trust trustworthy? *Social Psychology Quarterly*, pages 62–70.
- [Miller, 2018] Miller, T. (2018). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*.
- [Misztal, 2013] Misztal, B. (2013). *Trust in Modern Societies: The Search for the Bases of Social Order*. John Wiley & Sons.
- [Mitra and Gilbert, 2014] Mitra, T. and Gilbert, E. (2014). The language that gets people to give: Phrases that predict success on kickstarter. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pages 49–61. ACM.
- [Moser et al., 2017] Moser, C., Resnick, P., and Schoenebeck, S. (2017). Community commerce: Facilitating trust in mom-to-mom sale groups on facebook. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pages 4344–4357. ACM.
- [Mukherjee et al., 2014] Mukherjee, A. N., Bhattacharyya, S., and Bera, R. (2014). Role of information technology in human resource management of sme: A study on the use of applicant tracking system. *IBMRD's Journal of Management & Research*, 3(1):1–22.
- [Munro et al., 2010] Munro, R., Bethard, S., Kuperman, V., Lai, V. T., Melnick, R., Potts, C., Schnoebelen, T., and Tily, H. (2010). Crowdsourcing and language studies: the new generation of linguistic data. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 122–130.
- [Murray et al., 2012] Murray, N., Marchesotti, L., and Perronnin, F. (2012). AVA: a large-scale database for aesthetic visual analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2408–2415. IEEE.

- [Naaman et al., 2019] Naaman, M., Levy, K., and Hancock, J. (2019). Ai-mediated communication: Definition, research agenda, and ethical considerations. Working Paper.
- [Nannestad, 2008] Nannestad, P. (2008). What have we learned about generalized trust, if anything? *Annual Review of Political Science*, 11:413–436.
- [Netzer et al., 2016] Netzer, O., Lemaire, A., and Herzenstein, M. (2016). When words sweat: Identifying signals for loan default in the text of loan applications. In *Advances in Consumer Research*. Association for Consumer Research.
- [Newman and Antin, 2016] Newman, R. and Antin, J. (2016). Building for trust: Insights from our efforts to distill the fuel for the sharing economy. <http://nerds.airbnb.com/building-for-trust>. (Accessed May 2016).
- [Nyhan, 2000] Nyhan, R. C. (2000). Changing the paradigm: Trust and its role in public sector organizations. *The American Review of Public Administration*, 30(1):87–109.
- [Ott et al., 2011] Ott, M., Choi, Y., Cardie, C., and Hancock, J. T. (2011). Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 309–319. Association for Computational Linguistics.
- [Paszke et al., 2017] Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. (2017). Automatic differentiation in pytorch. In *Proceedings of the Workshop on Neural Information Processing Systems*.
- [Pavlou and Dimoka, 2006] Pavlou, P. A. and Dimoka, A. (2006). The nature and role of feedback text comments in online marketplaces: Implications for trust building, price premiums, and seller differentiation. *Information Systems Research*, 17(4):392–414.
- [Paxton, 2007] Paxton, P. (2007). Association memberships and generalized trust: A multilevel model across 31 countries. *Social Forces*, 86(1):47–76.
- [Pennebaker et al., 2001] Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001). Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.
- [Pennycook and Rand, 2019] Pennycook, G. and Rand, D. G. (2019). Fighting

- misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences*, 116(7):2521–2526.
- [Perez, 2018] Perez, S. (2018). Facebook is launching a new groups tab and plug-in. <https://techcrunch.com/2018/05/01/facebook-is-launching-a-new-groups-tab-and-plugin>. (Accessed Sep 2018).
- [Pew, 2007] Pew (2007). Americans and social trust: Who, where and why. <http://www.pewsocialtrends.org/2007/02/22/americans-and-social-trust-who-where-and-why>. (Accessed Sep 2018).
- [Pew, 2019] Pew (2019). Public trust in government: 1958–2019. <https://www.people-press.org/2019/04/11/public-trust-in-government-1958-2019/>. (Accessed Jul 2019).
- [Popat et al., 2016] Popat, K., Mukherjee, S., Strötgen, J., and Weikum, G. (2016). Credibility assessment of textual claims on the web. In *Proceedings of the ACM International Conference on Information and Knowledge Management*, pages 2173–2178. ACM.
- [Preece and Maloney-Krichmar, 2005] Preece, J. and Maloney-Krichmar, D. (2005). Online communities: Design, theory, and practice. *Journal of Computer-Mediated Communication*, 10(4).
- [Putnam, 1993] Putnam, R. D. (1993). The prosperous community. *The American Prospect*, 4(13):35–42.
- [Putnam, 1995] Putnam, R. D. (1995). Bowling alone: America’s declining social capital. *Journal of Democracy*, 6(1):65–78.
- [Putnam, 2000] Putnam, R. D. (2000). *Bowling Alone: The Collapse and Revival of American Democracy*. Simon and Schuster Nova York.
- [Qiu et al., 2018] Qiu, W., Parigi, P., and Abrahao, B. (2018). More stars or more reviews? In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, page 153. ACM.
- [Rainie and Wellman, 2012] Rainie, L. and Wellman, B. (2012). *Networked: The New Social Operating System*. Mit Press.

- [Rashkin et al., 2017] Rashkin, H., Choi, E., Jang, J. Y., Volkova, S., and Choi, Y. (2017). Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 2931–2937.
- [Reinecke et al., 2013] Reinecke, K., Yeh, T., Miratrix, L., Mardiko, R., Zhao, Y., Liu, J., and Gajos, K. Z. (2013). Predicting users’ first impressions of website aesthetics with a quantification of perceived visual complexity and colorfulness. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pages 2049–2058. ACM.
- [Ren et al., 2007] Ren, Y., Kraut, R., and Kiesler, S. (2007). Applying common identity and bond theory to design of online communities. *Organization Studies*, 28(3):377–408.
- [Resnick et al., 2000] Resnick, P., Kuwabara, K., Zeckhauser, R., and Friedman, E. (2000). Reputation systems. *Communications of the ACM*, 43(12):45–48.
- [Resnick and Zeckhauser, 2002] Resnick, P. and Zeckhauser, R. (2002). Trust among strangers in internet transactions: Empirical analysis of ebay’s reputation system. In *The Economics of the Internet and E-commerce*, pages 127–157. Emerald Group Publishing Limited.
- [Resnick et al., 2006] Resnick, P., Zeckhauser, R., Swanson, J., and Lockwood, K. (2006). The value of reputation on ebay: A controlled experiment. *Experimental Economics*, 9(2):79–101.
- [Ridings et al., 2002] Ridings, C. M., Gefen, D., and Arinze, B. (2002). Some antecedents and effects of trust in virtual communities. *The Journal of Strategic Information Systems*, 11(3-4):271–295.
- [Rini, 2019] Rini, R. (2019). Deepfakes are coming. we can no longer believe what we see. <https://www.nytimes.com/2019/06/10/opinion/deepfake-pelosi-video.html>. (Accessed Jul 2019).
- [Ritzer, 1975] Ritzer, G. (1975). Sociology: A multiple paradigm science. *The American Sociologist*, pages 156–167.
- [Rosenblat, 2018] Rosenblat, A. (2018). *Uberland: How Algorithms Are Rewriting the Rules of Work*. University of California Press.
- [Rosenblat and Stark, 2016] Rosenblat, A. and Stark, L. (2016). Algorithmic labor

and information asymmetries: A case study of uber’s drivers. *International Journal of Communication*.

- [Rother et al., 2004] Rother, C., Kolmogorov, V., and Blake, A. (2004). Grabcut: Interactive foreground extraction using iterated graph cuts. In *Proceedings of the ACM Transactions on Graphics*, volume 23, pages 309–314. ACM.
- [Rotter, 1967] Rotter, J. B. (1967). A new scale for the measurement of interpersonal trust. *Journal of Personality*, 35(4):651–665.
- [Rotter, 1971] Rotter, J. B. (1971). Generalized expectancies for interpersonal trust. *American Psychologist*, 26(5):443.
- [Rousseau and Greller, 1994] Rousseau, D. M. and Greller, M. M. (1994). Human resource practices: Administrative contract makers. *Human Resource Management*, 33(3):385–401.
- [Rubin, 1975] Rubin, Z. (1975). Disclosing oneself to a stranger: Reciprocity and its limits. *Journal of Experimental Social Psychology*, 11(3):233–260.
- [Schoorman et al., 2007] Schoorman, F. D., Mayer, R. C., and Davis, J. H. (2007). An integrative model of organizational trust: Past, present, and future. *Academy of Management Review*.
- [Schwarz et al., 2016] Schwarz, K., Wieschollek, P., and Lensch, H. (2016). Will people like your image? learning the aesthetic space. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*.
- [Shao et al., 2018] Shao, C., Ciampaglia, G. L., Varol, O., Yang, K.-C., Flammini, A., and Menczer, F. (2018). The spread of low-credibility content by social bots. *Nature Communications*, 9(1):4787.
- [Shu et al., 2019a] Shu, K., Cui, L., Wang, S., Lee, D., and Liu, H. (2019a). Defend: Explainable fake news detection. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.
- [Shu et al., 2017] Shu, K., Sliva, A., Wang, S., Tang, J., and Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1):22–36.
- [Shu et al., 2019b] Shu, K., Wang, S., and Liu, H. (2019b). Beyond news contents:

- The role of social context for fake news detection. In *Proceedings of the ACM International Conference on Web Search and Data Mining*, pages 312–320. ACM.
- [Shu et al., 2018] Shu, K., Zhou, X., Wang, S., Zafarani, R., and Liu, H. (2018). Understanding user profiles on social media for fake news detection. In *Proceedings of the IEEE Conference on Multimedia Information Processing and Retrieval*.
- [Silverman et al., 2018] Silverman, C., Lytvynenko, J., and Thuy Vo, L. (2018). How facebook groups are being exploited to spread misinformation, plan harassment, and radicalize people. <https://www.buzzfeednews.com/article/craigsilverman/how-facebook-groups-are-being-exploited-to-spread>. (Accessed Jul 2019).
- [Simon and Brown, 1987] Simon, B. and Brown, R. (1987). Perceived intragroup homogeneity in minority-majority contexts. *Journal of Personality and Social Psychology*.
- [Spence, 2002] Spence, M. (2002). Signaling in retrospect and the informational structure of markets. *The American Economic Review*, 92(3):434–459.
- [State et al., 2016] State, B., Abrahao, B. D., and Cook, K. (2016). Power imbalance and rating systems. In *Proceedings of the International AAAI Conference on Web and Social Media*, pages 368–377.
- [Stokes, 1983] Stokes, J. P. (1983). Components of group cohesion: Intermember attraction, instrumental value, and risk taking. *Small Group Behavior*.
- [Sun, 2010] Sun, H. (2010). Sellers’ trust and continued use of online marketplaces. *Journal of the Association for Information Systems*, 11(4):2.
- [Sundar, 1998] Sundar, S. S. (1998). Effect of source attribution on perception of online news stories. *Journalism & Mass Communication Quarterly*, 75(1):55–68.
- [Susser et al., 2018] Susser, D., Roessler, B., and Nissenbaum, H. (2018). Online manipulation: Hidden influences in a digital world. Working Paper.
- [Szegedy et al., 2016] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826.

- [Tan et al., 2014] Tan, C., Lee, L., and Pang, B. (2014). The effect of wording on message propagation: Topic- and author-controlled natural experiments on twitter. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.
- [Tavakolifard and Almeroth, 2012] Tavakolifard, M. and Almeroth, K. C. (2012). Social computing: an intersection of recommender systems, trust/reputation systems, and social networks. *IEEE Network*, 26(4):53–58.
- [Taylor et al., 2007] Taylor, P., Funk, C., and Clark, A. (2007). Americans and social trust: Who, where and why. *A Social Trends Report*.
- [Thebault-Spieker et al., 2017] Thebault-Spieker, J., Terveen, L., and Hecht, B. (2017). Toward a geographic understanding of the sharing economy: Systemic biases in uberx and taskrabbit. *ACM Transactions on Computer-Human Interaction*, 24(3):21.
- [Thebault-Spieker et al., 2015] Thebault-Spieker, J., Terveen, L. G., and Hecht, B. (2015). Avoiding the south side and the suburbs: The geography of mobile crowdsourcing markets. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work & Social Computing, CSCW '15*, pages 265–275. ACM.
- [Toma and Hancock, 2012] Toma, C. L. and Hancock, J. T. (2012). What lies beneath: The linguistic traces of deception in online dating profiles. *Journal of Communication*, 62(1):78–97.
- [Toma et al., 2008] Toma, C. L., Hancock, J. T., and Ellison, N. B. (2008). Separating fact from fiction: An examination of deceptive self-presentation in online dating profiles. *Personality and Social Psychology Bulletin*, 34(8):1023–1036.
- [Tully et al., 2019] Tully, M., Vraga, E. K., and Bode, L. (2019). Designing and testing news literacy messages for social media. *Mass Communication and Society*, pages 1–25.
- [Ugander et al., 2012] Ugander, J., Backstrom, L., Marlow, C., and Kleinberg, J. (2012). Structural diversity in social contagion. *Proceedings of the National Academy of Sciences*, pages 5962–5966.
- [Uski and Lampinen, 2014] Uski, S. and Lampinen, A. (2014). Social norms and self-presentation on social network sites: Profile work in action. *New Media & Society*.

- [Uzzi, 1996] Uzzi, B. (1996). The sources and consequences of embeddedness for the economic performance of organizations: The network effect. *American Sociological Review*, pages 674–698.
- [Vaidya et al., 2019] Vaidya, T., Votipka, D., Mazurek, M. L., and Sherr, M. (2019). Does being verified make you more credible? account verification’s effect on tweet credibility. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*. ACM.
- [Van Vugt and Hart, 2004] Van Vugt, M. and Hart, C. M. (2004). Social identity as social glue: the origins of group loyalty. *Journal of Personality and Social Psychology*, 86(4):585.
- [Vasilescu et al., 2012] Vasilescu, B., Capiluppi, A., and Serebrenik, A. (2012). Gender, representation and online participation: A quantitative study of stackoverflow. In *Proceedings of the International Conference on Social Informatics*, pages 332–338. IEEE.
- [Vigoda-Gadot and Talmud, 2010] Vigoda-Gadot, E. and Talmud, I. (2010). Organizational politics and job outcomes: The moderating effect of trust and social support. *Journal of Applied Social Psychology*, 40(11):2829–2861.
- [Vosoughi et al., 2018] Vosoughi, S., Roy, D., and Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380):1146–1151.
- [Walther and Bunz, 2005] Walther, J. B. and Bunz, U. (2005). The rules of virtual groups: Trust, liking, and performance in computer-mediated communication. *Journal of Communication*, 55(4):828–846.
- [Wang, 2019] Wang, P. (2019). This person does not exist. <http://thispersondoesnotexist.com>. (Accessed Jul 2019).
- [Wang et al., 2016] Wang, X., Sun, Z., Zhang, W., Zhou, Y., and Jiang, Y.-G. (2016). Matching user photos to online products with robust deep features. In *Proceedings of the ACM International Conference on Multimedia Retrieval*, pages 7–14. ACM.
- [Wilber et al., 2017] Wilber, M. J., Fang, C., Jin, H., Hertzmann, A., Collomosse, J., and Belongie, S. (2017). Bam! the behance artistic media dataset for recognition beyond photography. In *Proceedings of the IEEE International Conference on Computer Vision*.

- [Wu et al., 2010] Wu, G., Hu, X., and Wu, Y. (2010). Effects of perceived interactivity, perceived web assurance and disposition to trust on initial online trust. *Journal of Computer-Mediated Communication*, 16(1):1–26.
- [WVS, 2018] WVS (2018). World values survey database. <http://www.worldvaluessurvey.org/WVSContents.jsp>. (Accessed Aug 2018).
- [Yakovleva et al., 2010] Yakovleva, M., Reilly, R. R., and Werko, R. (2010). Why do we trust? moving beyond individual to dyadic perceptions. *Journal of Applied Psychology*, 95(1):79.
- [Yamagishi and Cook, 1993] Yamagishi, T. and Cook, K. S. (1993). Generalized exchange and social dilemmas. *Social Psychology Quarterly*, pages 235–248.
- [Yamagishi et al., 2009] Yamagishi, T., Matsuda, M., Yoshikai, N., Takahashi, H., and Usui, Y. (2009). Solving the lemons problem with reputation. In *eTrust: Forming Relationships in the Online World*.
- [Yamagishi and Yamagishi, 1994] Yamagishi, T. and Yamagishi, M. (1994). Trust and commitment in the united states and japan. *Motivation and Emotion*, 18(2):129–166.
- [Yao et al., 2017a] Yao, W., Dai, Z., Huang, R., and Caverlee, J. (2017a). Online deception detection refueled by real world data collection. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*.
- [Yao et al., 2017b] Yao, Y., Viswanath, B., Cryan, J., Zheng, H., and Zhao, B. Y. (2017b). Automated crowdturfing attacks and defenses in online review systems. In *Proceedings of the ACM Conference on Computer and Communications Security*.
- [Yu and Grauman, 2014] Yu, A. and Grauman, K. (2014). Fine-grained visual comparisons with local learning. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*.
- [Yu and Grauman, 2017] Yu, A. and Grauman, K. (2017). Semantic jitter: Dense supervision for visual comparisons via synthetic images. In *Proceedings of the International Conference on Computer Vision*.
- [Yuki et al., 2005] Yuki, M., Maddux, W. W., Brewer, M. B., and Takemura, K. (2005). Cross-cultural differences in relationship-and group-based trust. *Personality and Social Psychology Bulletin*, 31(1):48–62.

- [Zelmer, 2003] Zelmer, J. (2003). Linear public goods experiments: A meta-analysis. *Experimental Economics*, 6(3):299–310.
- [Zervas et al., 2015] Zervas, G., Proserpio, D., and Byers, J. (2015). A first look at online reputation on airbnb, where every stay is above average. Working Paper.
- [Zhang et al., 2018] Zhang, A. X., Ranganathan, A., Metz, S. E., Appling, S., Sehat, C. M., Gilmore, N., Adams, N. B., Vincent, E., Lee, J., Robbins, M., et al. (2018). A structured response to misinformation: Defining and annotating credibility indicators in news articles. In *Proceedings of the International Conference on World Wide Web*, pages 603–612. International World Wide Web Conferences Steering Committee.
- [Zhang et al., 2007] Zhang, J., Ackerman, M. S., and Adamic, L. (2007). Expertise networks in online communities: structure and algorithms. In *Proceedings of the International Conference on World Wide Web*, pages 221–230. ACM.
- [Zhao et al., 2012] Zhao, L., Lu, Y., Wang, B., Chau, P. Y., and Zhang, L. (2012). Cultivating the sense of belonging and motivating user participation in virtual communities: A social capital perspective. *International Journal of Information Management*, 32(6):574–588.
- [Zheng et al., 2009] Zheng, X. S., Chakraborty, I., Lin, J. J.-W., and Rauschenberger, R. (2009). Correlating low-level image statistics with users-rapid aesthetic and affective judgments of web pages. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pages 1–10. ACM.
- [Zhou et al., 2008] Zhou, Y., Wilkinson, D., Schreiber, R., and Pan, R. (2008). Large-scale parallel collaborative filtering for the netflix prize. In *International conference on algorithmic applications in management*, pages 337–348. Springer.