Synthesized Social Signals: Computationally-Derived Social Signals from Account Histories

Jane Im[†], **Sonali Tandon**[†], **Eshwar Chandrasekharan**[††], **Taylor Denby**[†], **Eric Gilbert**[†]
[†]University of Michigan, [††]Georgia Institute of Technology
[†]Ann Arbor MI USA, [††]Atlanta GA USA
imjane@umich.edu, tsonali@umich.edu, eshwar3@gatech.edu, tdenby@umich.edu, eegg@umich.edu

ABSTRACT
Social signals are crucial when we decide if we want to interact with someone online. However, social signals are typically limited to the few that platform designers provide, and most can be easily manipulated. In this paper, we propose a new idea called *synthesized social signals (S3s)*: social signals computationally derived from an account's history, and then rendered into the profile. Unlike conventional social signals such as profile bios, S3s use computational summarization to reduce receiver costs and raise the cost of faking signals. To demonstrate and explore the concept, we built *Sig*, an extensible Chrome extension that computes and visualizes S3s. After a formative study, we conducted a field deployment of Sig on Twitter, targeting two well-known problems on social media: toxic accounts and misinformation. Results show that Sig reduced receiver costs, added important signals beyond conventionally available ones, and that a few users felt safer using Twitter as a result. We conclude by reflecting on the opportunities and challenges S3s provide for augmenting interaction on social platforms.

CCS Concepts
•Human-centered computing → Collaborative and social computing systems and tools;

Author Keywords
social computing; social signals; social media; social platform

INTRODUCTION
Discovering new people—and deciding whether we want to interact with them—is a fundamental experience online [13]. Interacting with new people online can bring us in touch with new information [18], new experiences [16], and new opportunities [34]. At the same time, interacting with someone new also brings risks: 40% of U.S. adults have been harassed online, and half of those did not know the perpetrator [15]. Nearly a quarter of U.S. adults have spread misinformation—with most first exposed to it by someone they did not already know [4]. Yet, deciding which is which (helpful vs. hurtful) is often challenging.

CHI '20, April 25–30, 2020, Honolulu, HI, USA.

ACM 978-1-4503-6708-0/20/04 ...$15.00.
http://dx.doi.org/10.1145/3313831.3376383

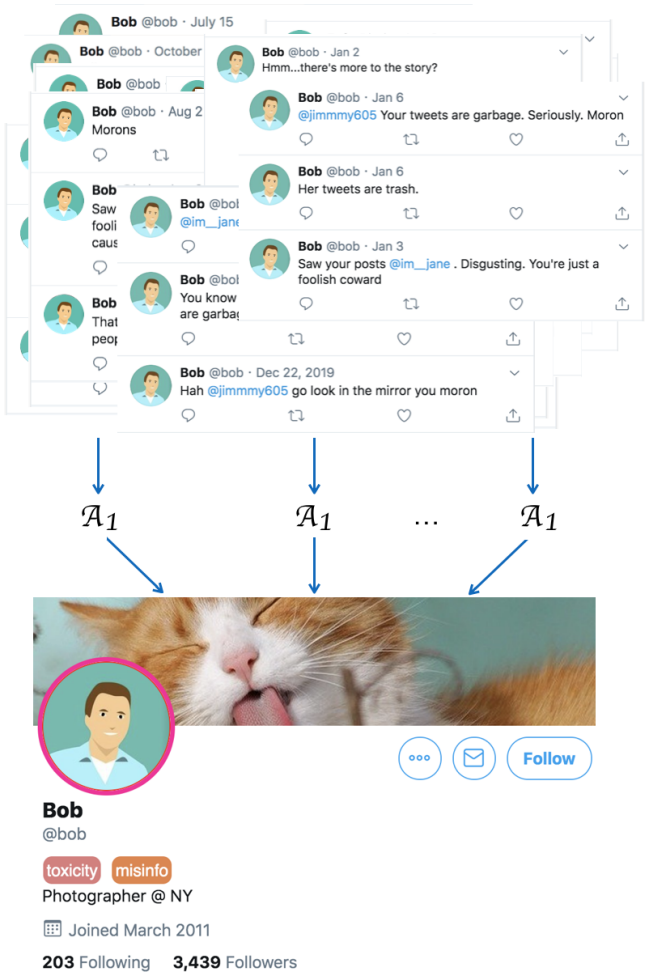

Figure 1. Illustration of *synthesized social signals (S3s)*. @bob's account history flows through different algorithms, A_1, A_2, ... A_n, to produce signals that are then rendered into the profile. @bob's profile has been augmented with signals corresponding to authoring toxic messages and spreading misinformation.

To make these decisions, we use *social signals* [13]. Online, social signals are typically features provided by platform designers that allow users to express themselves. They include profile images, bios, location fields, cover images, etc.—even the most subtle of which we use to form quick impressions of people [16]. More formally, signals are "perceivable features and actions that indicate the presence of hidden qualities" [13], and are the primary inputs to social cognition [3].

However, unlike their face-to-face (f2f) counterparts (e.g., fashion, hairstyles, etc.), online social signals are limited to the few platform designers explicitly provide [25, 45]. In other words, there are comparatively fewer cues about someone online than there would be f2f. Perhaps more importantly, online social signals are generally easier to fake than their offline counterparts. For example, someone might buy a pair of non-prescription glasses to signal intelligence and sophistication before a date; on an online dating site, they can present a completely new face with a 5-second image search. Scholars often refer to these phenomena in terms of the *cost* of faking a signal (i.e., the cost of buying the glasses vs. the cost of performing the image search) [13].

Despite this impoverished set of cues online, and the ease of faking them, online social platforms do have a largely untapped resource for understanding people online: *post histories*. When someone new drops by your office for a chat, they don't bring with them a history of every interaction they've ever had before. Yet, profile pages provide exactly this—and we argue—could be used to *augment* online social interaction beyond f2f metaphors [26]. However, to make use of the data archived in account histories, a user would have to manually read through each one.[1] At the scale of thousands of posts per account, the *receiver cost* of gathering relevant signals—the cost borne by the person trying to assess the signal—is high.

To bridge this gap, we introduce a novel computational approach called *synthesized social signals (S3s)* (see Figure 1). Synthesized social signals are signals computationally derived from a user's history using algorithms, and then rendered into the profile. Computational summarization obviates the need for people to scan every single post. That is, S3s aim to significantly reduce receiver costs with algorithms. Moreover, because they rely on deep archives of profile data, the costs related to deceiving with them are higher. In other words, someone would have to invest a significant amount of effort to delete or modify large portions of their posts in order to effectively manipulate S3s.

To demonstrate and explore S3s, we built a new system called *Sig*. Sig is an extensible Chrome extension that computes and visualizes S3s in the profile. After a survey-based formative study with 60 people, and an iterative design process, we conducted a field deployment of Sig on Twitter. The deployment targeted two well-known problems on social media: toxic accounts and misinformation [15, 35]. During a multi-day field study with 11 users recruited online, Sig computed toxicity and misinformation S3s on accounts in real-time—with its renderings visible on the Twitter timeline, notification page, and profile pages. The field study was followed by interviews and a survey.

Participants used Sig's S3s to mute, block, and unfollow accounts. In one case, Sig's S3s contravened a conventional signal (the Twitter blue checkmark) with a misinformation S3, which led them to see a supposedly trustworthy account in a different light. Broadly speaking, Sig's S3s reduced receiver costs, and some users reported feeling safer as a result.

This paper makes three contributions. First, we introduce a conceptual overview of *synthesized social signals*—social signals computed from a user's history of posts using algorithms, and then rendered in the profile. Second, we present a system contribution in the form of Sig, an extensible Chrome extension that generates S3s on social media platforms. Finally, we present the results of formative and field studies that illustrate the opportunities and challenges S3s present for augmenting interaction on social platforms.

RELATED WORK

Here we review past literature on social signals' effects on social perception on social platforms, as well as signaling theory. Next, we review prior work on system-generated cues in social platforms.

Importance of Social Signals in Social Platforms

People manage their self-presentation strategies on social platforms using various signals, such as crafting bio messages in order to balance accuracy and desirability [16]. Past research has shown that such self-presentation through social signals profoundly affects people's perception of one another. For instance, profile images impact people's impressions and decisions of other people. Negative facial expressions and the absence of profile images negatively impact a buyer's interest in using peer-to-peer accommodation rental services [20], as well as how people react to bystander intervention [31]. People also attend to small cues in bio messages when choosing dating partners [17] and use information in profile fields (e.g., school name, birthday, interests) when connecting with people because they reduce cost in finding common referents [30]. These social signals are all conventional signals, which are generally easier to fake [13, 23]. Nevertheless as prior research shows, conventional social signals greatly influence people's judgements, as social platforms are designed around them [17, 20, 31]. Other social signals that are relatively more reliable are often costlier to access. For instance, a public list of friends is more reliable in that it provides verification of identity claims (because other people essentially are involved in producing the list) [8, 13, 14]. However, it costs more time and effort to peruse a friend list.

This work. We extend this line of work from a design point of view: we introduce a new way to design cost-efficient, reliable social signals using cues left in an account's history of posts.

Cost in Signaling

Cost is an important factor in both sending and receiving signals. According to signaling theory, which stemmed from both biology [11, 21, 38, 50] and economics [39, 44], receivers value costly signals, perceiving them as more reliable [37]. This is a motivation to send a costly signal (to get better responses) and also causes people to try to fake costly signals (which is hard, by definition). However, a problem here is that it takes *cost to evaluate* such reliable signals as well [13]. High receiver costs cause people to end up relying on social signals that are easily accessible but less reliable [11, 51]. The same

[1] During our formative survey, introduced shortly, we found that participants did this, and viewed account histories as the most important piece of information when encountering someone new on Twitter (see Table 1).

applies to social platforms. People put importance on and respond to social signals that are costly to make [46, 47]. High receiver costs however cause people to rely on conventional social signals that are easy to access but easy to manipulate such as profile bios [13].

This work. S3s aim to reduce receiver cost for deriving information from accounts' post history using computation. At the same time, we argue it is hard to fake S3s since they are computational summaries of many posts over time.

System-generated Cues in Social Platforms

Who produces the cues also varies the cues' utility. User-generated cues are ones the user directly inputs into the system, which the system then shows verbatim; system-generated cues (of which S3s are a type) process or aggregate information and then render the signal [41]. An example of the latter is the "friend list," which represents an aggregation database query. The warranting value of system-generated cues is generally higher in social platforms [41, 42]. It is comparatively costly to manipulate system-generated cues, although not impossible. For instance, the number of friends is a well-known system-generated cue [1]. Although a person has some control over how many friends they have (for example, a nefarious user could buy bots to follow them on a gray market [6]), the person cannot choose not to show the number at all. The system is configured to show it.

Compared to user-generated cues however, system-generated cues remain relatively unexplored. Computational approaches to synthesize various signals have been explored by prior work such as bot detection [10]. But *how and where to render this information as accessible information* so that end-users can use it remains as an open question. An example of building new system-generated cues is Comment Flow, which aimed to provide easy access to information of interactions within networks [32]. Comment Flow can be seen as building a new system-generated cue out of the interaction between accounts and reducing the cost of manually tracing friends of an account [13, 32]. Other systems have focused on deriving account-level information on social media platforms as well. Seriously Rapid Source Review (SRSR) helps journalists filter and assess information sources on Twitter via account-level characteristics [12]. Botometer computes the likelihood whether a given Twitter account is a bot based on features including content and language of tweets [43]. Compared to SRSR, Comment Flow, and Sig however, Botometer's intent is different in that it focuses on classification results instead of rendering the information in situ.

This work. In this paper, we build on the line of work of system-generated cues by introducing a new extensible, browser-based framework for S3s to be rendered in profiles. In short, Sig works as a gateway for *rendering* classifications of systems like Botometer, empowering users to make decisions based on them in real-time interaction.

SYNTHESIZED SOCIAL SIGNALS (S3S)

As explained above, social signals have profound effects on on social behavior (e.g., [30, 31]). Figure 2 illustrates some of the myriad social signals available to a viewer of a profile on

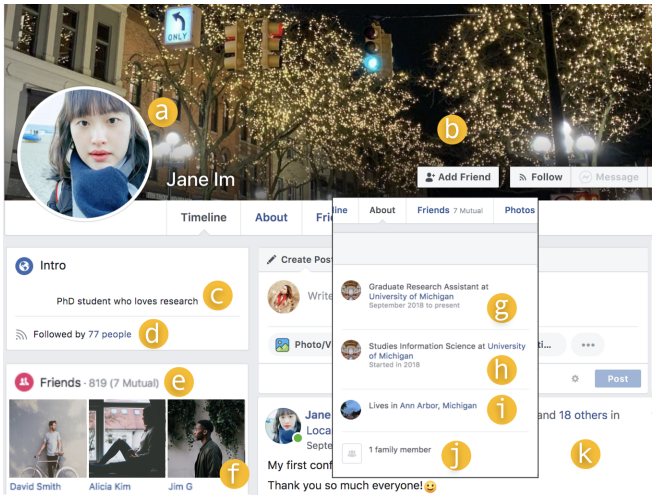

Figure 2. Example social signals on Facebook: a) profile image, b) cover image, c) bio, d) # of followers, e) # of friends (# of mutual friends), f) public list of friends, g) work information, h) school information, i) location information, j) family member information. Past posts in profile page k) contain cues from which we derive S3s.

Facebook. Many tens of social signals exist to provide context about the person represented: profile image, bio, number of friends, work history, location, number of mutual friends, etc. However, all of these (with the notable exception of system-generated signals such as number of friends) were typed in by the user. A motivated actor can fake these if they so choose [23, 36, 37].

Definition

In this paper, we argue that systems could make use of modern algorithms to render new social signals from entire account histories. We call these *synthesized social signals*, or *S3s*. An algorithm, perhaps making use of modern machine learning methods, consumes a history of post behavior, and converts it into usable S3s, which are then rendered in the profile. S3s live in the profile so as to put social information in the place where it matters most for social cognition. Unlike the conventional social signals in Figure 2, S3s derived in this way would have unique cost and reliabilty attributes.

Cost and Reliability

As reviewed earlier, a central part of social signaling is the reliability of signals. This is inextricably linked with the cost associated with producing it: higher cost signals have higher reliabilty. As Donath writes, there is a difference between lifting weights for a year to become strong and buying a "Gold's Gym" t-shirt at the store [13]. While they both aim to convey the social signal of *strength*, the former is costly and reliable, while the latter is cheap and unreliable.

Faking S3s has high cost. While they are certainly not unassailable, since S3s derive from computational summaries of large numbers of posts over time, a motivated actor trying to fake them would have to manipulate many, many posts to fool them. This is not impossible; it is just costly. Therefore, in the language of Donath, S3s have high reliability.

Rank	Social Signal	Mean (σ)
1	Past tweets & replies of the account	4.32 (0.94)
2	Past tweets from others that the account replied to	3.95 (1.01)
3	Profile description	3.90 (1.14)
4	Past media used by the account	3.70 (1.11)
5	Past likes of the account	3.50 (1.20)
6	Profile image	3.42 (1.19)
7	List of followers	3.30 (1.44)
8	Number of followers	3.29 (1.49)
9	Number of tweets	3.26 (1.32)
10	Website	3.22 (1.43)
11	List of following	3.19 (1.34)
12	Cover image	3.03 (1.18)
13	Number of following	3.02 (1.38)
14	Number of likes	2.97 (1.55)
15	Joined date	2.88 (1.47)
16	Location	2.84 (1.44)

Table 1. Result of non-probability survey asking participants to indicate the importance they put on various social signals when deciding to whether interact with a stranger account on Twitter (all questions are on 1-5 Likert scale).

Another central aspect of social signaling is the cost borne by the person interpreting the signal. This is called the receiver cost [13, 22]. In Figure 2, the profile image can be quickly and easily understood by a viewer. It takes little cognitive energy, as we are hard-wired to focus on and process faces [2]. On the other hand, arriving at a gestalt of a person from their history of posts would involve time spent reading each one, perhaps its surrounding context, the relationships involved, etc. This manual process would have very high receiver costs associated with it.

Algorithms, while inherently imperfect approximations of our own abilities to summarize human behavior, can act as helpful tools. By summarizing data that a person would struggle to process on their own, S3s can dramatically reduce receiver costs, while at the same time increasing costs for those would want to deceive through them.

FORMATIVE STUDY

Method

In this paper, we introduce a new system built on the concept of S3s called Sig. First, we discuss the formative work done to drive its design. We conducted an online survey to: 1) get insights into which signals Twitter users thought were important when interacting with strangers; and, 2) get feedback on an early prototype of Sig, our system for generating and visualization S3s. A demo of Sig was recorded as a video and linked to the survey. We tweeted about our survey and recruited participants on Twitter by promoting the tweet through Twitter Ads. All participants were given $10 Amazon gift cards as compensation. Among 67 responses, 7 responses were discarded as they were duplicates. 33 identified as women, 23 identified as men, 2 identified as non-binary, and 2 did not disclose their gender. Age ranged from 18 to 55 with an average age of 28 (σ=8.1). Thirty-six identified as white, 11 identified as Asian, 9 identified as Hispanic or Latino/a/x, 1 identified as American Indian, and 1 identified as Black and White, and 2 did not disclose their race. The first author coded the answers to the open-ended questions using an inductive coding approach. Themes were then discussed with other authors.

Findings

What people look for when interacting with strangers

When asked to rate the importance of signals out of "Very important (5)" to "Very unimportant (1)", *past tweets and replies of an account* received the highest average score out of all signals (Table 1). As a non-probability sample, care should be taken generalizing from these results. However, as a formative survey this supplied valuable feedback about moving forward.

Open-ended feedback on the preliminary system

The majority of participants replied they thought the tool would be helpful in interacting with strangers on Twitter.

> *"Yes! It's just a quick heads up before you enter any Internet shenanigans."*

> *"Manually going through signals takes time and energy. Having it easily available helps.".*

Several survey participants noted they wanted higher transparency from the system. Many participants said they would like to know how we defined 'toxicity' of an account. Thus we made it clear in the interface that we were using Perspective API[2] and OpenSources[3] when determining the toxicity and misinformation-spreading behavior of accounts. We also introduced a modal that shows flagged tweets as shown in Figure 4, which many survey participants liked because it showed the reason behind the flagging.

THE SYSTEM: SIG

To demonstrate and explore the concept of S3s, we built *Sig*. Sig is an extensible Chrome extension that computes and visualizes S3s on social platforms, and then renders them into profiles. While Sig can be run on multiple platforms, we will focus on its application to Twitter, the site of our field deployment. Here we describe the features as well as the algorithms used for deriving Sig's S3s.

User Scenarios

Next, two scenarios illustrate how Sig can be used. Both are derived from actual field study participant experiences.

Scenario A. One of Michelle's followers retweeted a tweet containing a meme. She thought the tweet was funny, but noticed Sig flagged the account that tweeted it. She goes to check the account profile and finds it is marked as toxic. By clicking on the "toxicity" tag, she discovers that many tweets were aggressive to others and included offensive racial slurs. *Glad Sig prevented me from following the account*, she thought while she quickly closed the profile page.

Scenario B. Among the accounts that Twitter recommends to follow ("Who to follow"), one account's profile picture catches Taylor's eye. Curious, he goes to the account's profile

[2] https://www.perspectiveapi.com/

[3] https://github.com/BigMcLargeHuge/opensources

Figure 3. Profile border color is used to render S3s in notification/timeline page of Twitter. A red border indicates that at least one S3 has been triggered, and the account may be risky to interact with (examples: Alice and John, in first and third row). A blue, double-lined border indicates that Sig is currently computing S3s for the account (examples on the right in the first row). If no S3s are triggered, the blue border disappears after computation (examples in second and right of third row).

page and sees that Sig flagged it as misinformation-spreading. Clicking the "misinformation" tag, he sees that the account has shared a lot of links from suspicious sources. *I for sure don't want to see this account again*, Taylor thought while he muted the account.

Sig Features: S3s in Sig

Sig computes and visualizes S3s. Currently for Twitter, the system functions on 1) profile pages, 2) the timeline, and 3) the notification page. We oriented Sig to support S3s which indicate potentially problematic behavior. It signals such behavior to users by encircling profile borders (see Figure 3). In part because the average latency for computing some S3s was about 10 seconds, Sig does not activate on accounts that a user follows on Twitter (we roughly assume that accounts a person already follows are "not strangers" and therefore less useful for Sig to operate on). When a user first logs into Sig, Sig locally stores a list of accounts that the user follows in order to ensure that it does not run on the accounts within that list.

Profile page. Whenever a user visits another account profile page, Sig computes S3s (here toxicity and misinformation-spreading behavior) by fetching up to 200 tweets[4] and running the algorithms on them. If at least one of an account's S3s is over the threshold the user has set, then the border of the profile page turns red to warn the user (Figure 1). In short, Sig is warning the user that they need to be careful when trying to decide whether or not to interact with the account because the past posting history shows the account is likely to be either toxic or misinformation spreading. Tags indicating which signal is above the threshold also appear above the bio. The user is able to expand details on each S3 by clicking each of the tags that appear (Figure 1).

[4]Due to Perspective API's latency (as Perspective API was run on each tweet) we took 200 as a tradeoff between depth and latency.

Figure 4. Modal showing up to five tweets of an account flagged for toxicity S3. Per the pilot study results, we aimed for ensuring transparency in how Sig shows S3s.

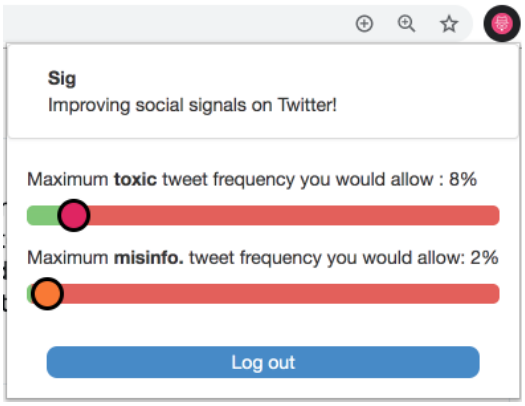

Figure 5. Users can adjust the sliders to set their own thresholds for each S3 in Sig. In the field deployment, participants could choose thresholds for toxicity and misinformation.

Notification and Timeline. Sig also shows signals on the notification page, as it is the main way users become aware of interactions with other accounts, such as liking, mentioning, and following. As shown in Figure 3, the profile's border indicates the current state of Sig. When the border is blue and double lined, it indicates that Sig is computing S3s for that account. When the border turns red, it means that there is at least one S3 triggered (e.g., either the account is likely to be toxic or spreading misinformation). If the blue border disappears and no further changes happen on the profile image, it indicates that Sig finished its computation. When the user is scrolling down in the timeline, S3s are computed in real time and visualized by coloring the profile image's borders. This enables users to quickly check on accounts while scrolling through the timeline and reading tweets.

Transparency behind the system's decision

Based on the survey and pilot study, participants wanted transparency behind the judgements of the extension. Thus, we added a modal showing the content of up to five tweets flagged for S3s when an account is flagged (Figure 4). For misinformation S3, we also highlight the URL that was categorized as misinformation spreading.

Features for Customization

Because people's tolerance for toxic and misinformation-spreading behavior might be different and machine learning models have inherent limitations, we let users choose their own thresholds for toxicity and misinformation-spreading behavior. When a user logs in, sliders appear for each S3 which users can use to set their own thresholds (Figure 5). Users can choose their own threshold anytime by clicking the small icon in the upper right side of the browser.

System Implementation

Sig consists of a Chrome browser extension and a server that handles the requests. When a user installs Sig, they authenticate through their Twitter account using OAuth[5]. This enables Sig to use each study participant's API keys to send requests to Twitter API when retrieving tweets.

The server is built using the Django framework, Gunicorn and Nginx. On average it took about 10 seconds per account to compute toxicity and misinformation S3s and render them on Twitter. The latency was mainly due to Perspective API's rate limit. Thus, the system caches the S3s of accounts already seen by the participant using MySQL database, as well as Chrome's localStorage. For the field study participant's privacy, we do not store which participant saw which account (just the account's S3s). So every time anyone sees or visits the same account they are be retrieved quickly. The same data is also cached into each participant's browser. Each participant's Twitter API keys and accounts that one follows were only cached in the localStorage and not in the MySQL database, for privacy reasons.

S3s' algorithms

Sig has an extensible framework for plugging arbitrary S3s. Here we outline the algorithms for the S3s used in the field deployment.

Toxicity. To compute each account's toxicity level, we score tweets made by the account using the Perspective API[6]. The Perspective API uses a Convolutional Neural Network (CNN) trained with word vector inputs from a corpus of Wikipedia discussion comments labeled to contain personal attacks [49]. Given a new comment, the Perspective API computes a score between 0 and 1 that indicates the likelihood that someone perceives the message as "toxic." We use three specific types of toxicity scores defined by the API as follows:

- *TOXICITY*: "rude, disrespectful, or unreasonable comment that is likely to make people leave a discussion."
- *SEVERE_TOXICITY*: "a very hateful, aggressive, disrespectful comment or otherwise very likely to make a user leave a discussion or give up on sharing their perspective. This model is much less sensitive to comments that include positive uses of curse words, for example."[7]
- *ATTACK_ON_COMMENTER*: "attack on the author of an article or post." This model was trained on New York Times data tagged by their moderation team.

Through our pilot study (described in the next section), we found that using $TOXICITY > 0.8$ alone flags tweets with curse words used in positive contexts. As many pilot-study participants did not consider these tweets to be undesirable, we made the toxicity criteria stricter to increase precision: $TOXICITY > 0.8$ *and* $SEVERE\_TOXICITY > 0.7$ *and* $ATTACK\_ON\_COMMENTER > 0.5$. We feed each tweet to Perspective API and retrieve TOXICITY, SEVERE_TOXICITY, and ATTACK_ON_COMMENTER scores. Then the system calculates the proportion of tweets by the author that satisfy all three conditions in our toxicity criteria. The percentage of tweets satisfying the toxicity criteria among all tweets made by a Twitter account becomes the final toxicity level of the account. In other words, an account's toxicity level indicates its proportion of toxic tweets.

Misinformation criteria. To compute misinformation-spreading behavior, we used the OpenSources dataset[8], an expert-curated resource for "assessing online information sources, available for public use" and used by previous work on misinformation [7, 27]. We use the *conspiracy*, *fake*, and *bias* categories, which generated 486 domain names. For each tweet that contains a link, Sig parses out the domain name and checks if it is included in the dataset. For each account, the percentage of tweets that include a link from one of the 486 domains names is the final score for misinformation.

ITERATION: PILOT FIELD STUDY

Next, we briefly describe a pilot study that influenced our final design and implementation decisions. We recruited 13 people from Twitter by tweeting about our study and using Twitter Ads to promote our tweet. Over the course of four days, the participants used an early version of Sig for at least 30 minutes per day. After four days, the participants were interviewed about their experience of using the extension. All participants were compensated $30 for their participation. The 13 participants had varied occupations, from university students to illustrator and web developer.

Implementation Changes

Many participants (9/13) expressed interest in using a refined version of the extension which further motivated our work. However, the pilot study revealed issues that were crucial to cover for the final study.

Issue of false positives

The main issue revealed during the pilot study was false positive accounts labeled as toxic; that is, accounts that were flagged as toxic but did not seem so to participants. Almost all participants mentioned when looking at the flagged tweets in the modal after clicking the "toxicity" tag (Figure 4), many flagged tweets seemed to be flagged due to curse words, which they did not necessarily think indicated an account is toxic. This led us to change the thresholds for flagging tweets as toxic in the final iteration (as described earlier in the previous

[5] https://oauth.io/

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

[7] Perspective API released the *SEVERE_TOXICITY* model particularly to reduce such false positives—*SEVERE_TOXICITY* is much less sensitive to comments that use of curse words.

[8] https://github.com/BigMcLargeHuge/opensources

	Occupation	# of followers	# of following	Joined Date
P1	College Student	150	400	2012
P2	College Student	200	700	2014
P3	Paralegal	300	300	2017
P4	College Student	400	600	2013
P5	College Student	2000	400	2014
P6	Delivery Driver	100	1300	2016
P7	Sales associate	200	2100	2009
P8	Business consultant	300	300	2016
P9	College Student	200	400	2014
P10	College Student	550	400	2013
P11	Scholar	100	100	2018

Table 2. Descriptive statistics about field deployment participants. Numbers are rounded to the nearest hundred for participants' privacy.

section). Furthermore, many participants thought Sig labeled flagged accounts as definitely toxic. In order to prevent this, we added reminders in the final study instructions as well as in the extension's interface that we are not guaranteeing the flagged accounts are toxic or misinformation sharing but *providing a heads-up of some likelihood that they might be*.

Transparency in system decisions

Despite the false positive issue described above, all participants said they liked the feature of being able to see the flagged tweets (even when it showed them false positive tweets). Participants mentioned that they liked being able to easily see flagged tweets as the system pulled them up quickly, and they were able to use their own judgement if they wanted. "*...at the end of the day you still use your own judgment, but it just helps you by bringing those tweets to the forefront, basically."* Thus, in the final study, we created a feature that shows up to the top 5 most toxic tweets (Figure 4).

Visualization of signals

Almost all participants expressed satisfaction with the visualization of the signals (12/13). Thus, we did not change the overall visualization except for adding a timestamp to the tweets, which one participant wanted. Some participants wanted to see a safety mark on accounts that are not flagged. However, we did not want the absence of Sig's S3s to imply safety, so we did not implement this feature.

EVALUATION: FIELD STUDY

Next, we describe the design and results of a multi-day field deployment in which participants used Sig in the context of their typical, everyday Twitter use.

Method

We recruited 11 people from Twitter by tweeting about our study. We again used Twitter Ads to promote the tweet about the study as we did not want participants to come only from within the authors' networks.[9] Over the course of four days, we asked 11 participants to use the newly refined version of Sig for at least 30 minutes per day, over at least four days. Participants were compensated $30.

[9]However, since the Twitter algorithm promoted the ad to people with backgrounds similar to the authors' backgrounds, many participants ended up being college students.

Participants were given instructions to use the extension, but were not told to go and actively look for toxic or misinformation-spreading accounts for clear ethical reasons. We also wanted them to use the extension in a realistic and natural setting. After using the extension, all participants completed a survey and participated in a 30-50 minute interview. All interviews were conducted remotely using voice or video calls. During an interview, we first asked about the participants' interaction with strangers on Twitter. Next, we asked questions about their experience using the tool. We asked the participants to tell us how they used Sig, whether they encountered any flagged accounts, and if so, asked them to recall the accounts and what they thought about them. We also asked if there are other social signals besides toxicity and misinformation they wanted to add to Sig. In short, we wanted the participants to think about S3s on platforms while using the extension. The interviews were later transcribed, and the first and fourth author used an inductive coding approach to develop themes, using Dedoose[10]. Later, the authors assessed major themes.

Participants

Table 2 shows the occupation as well as number of followers, number of following, and joined date of the 11 participants. Seven participants identified as women, two identified as men, and two identified as non-binary. Five participants identified as white, two identified as Black or African-American. Two participants identified as Asian, one as Hispanic or Latino/a/x, and one participant identified as White and Asian. Age ranged from 18 to 33, with the average age being 23.

Results

We next report results from the interviews and survey. Ten participants (10/11) said they had seen accounts either flagged for toxic or misinformation during the study while scrolling through the timeline or going through profile pages of accounts that either tweeted with trending hashtags, participated in a thread, popped up in the Explore page[11], or were recommended by Twitter (10 participants saw accounts flagged as toxic and 2 participants saw accounts flagged as spreading misinformation). All 10 participants that saw flagged accounts confirmed Sig's S3s: they thought many flagged accounts were legitimately toxic or misinformation-spreading.

Overall experience using Sig. Our interview and exit survey results show that overall the participants were positive about their experience of using Sig. As shown in Figure 6, participants rated their overall experience of using Sig an average of 4 out 5, while the interviews revealed that many participants (10/11) were positive about Sig. The one participant (P8) that did not think Sig was helpful said that she was indifferent to interacting with strangers in the first place. Many participants said they liked that Sig gave a "heads-up" on the account's likely characteristic based on its history.

> "*I didn't have any moments where I said, 'I wish it was able to do this,' because it was doing what I wanted it*

[10]https://www.dedoose.com/

[11]https://twitter.com/explore

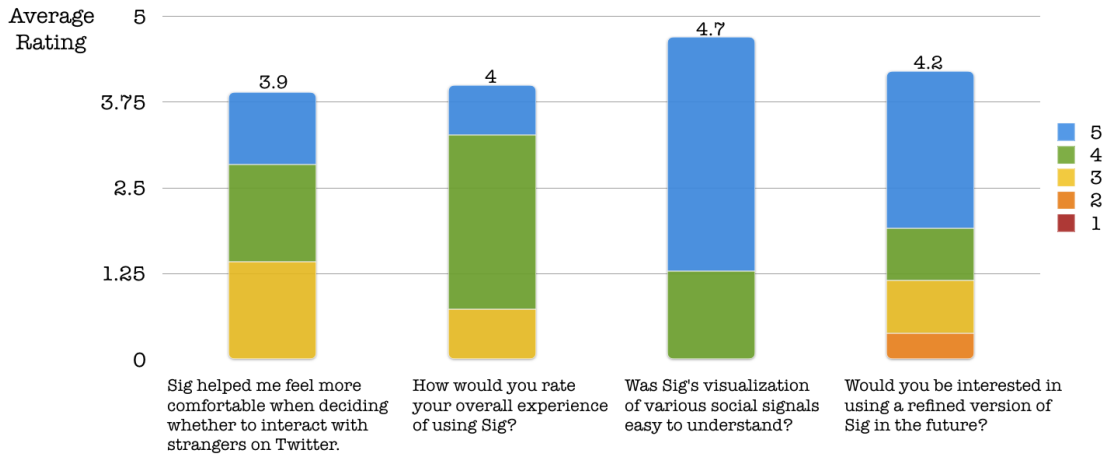

Figure 6. Results of the survey taken by participants after the field study. Each question appeared on a 1-5 Likert scale.

this entire time. It found if an account had a history of posting of toxic posts." –P6

"*I liked the fact that it would give me that information up front.*" –P1

"*...it just seems like most people feel like they're entitled to say whatever they want...If they encounter one, they'd just block and move on and they don't necessarily need an app to do that.*" –P8

Augmenting social decision-making. Despite the fact that the authors had not encouraged participants to use Sig to follow, mute, or block accounts (in order to not bias them before the study), many participants (6/11) reported that they used Sig to make social decisions on stranger accounts—they decided whether or not to follow, mute, or even block stranger accounts they encountered during the study. The six participants mentioned they easily noticed the circles or tags rendered by Sig (Figures 1 and 3) and then checked the modal to see the flagged tweets (Figure 4). If they thought the flagged account was legitimately toxic or misinformation-spreading, they muted, blocked, or decided not to follow the account. For instance, P6 said Sig flagged a somewhat ordinary looking account as misinformation-spreading—soon after, P6 realized it was sharing links to non-accredited news sites. After going through the profile page, P6 decided to block the account.

"*It wasn't a big name, a speaker, or a political activist, or anything like that. It was just someone on Twitter. [...] They were sharing links to non-accredited news sites. [...] The one that I mentioned earlier that flagged misinformation, blocked that one.*" –P6

"*Double check to see if the flagged tweets matched up with what I thought was a problem, which it usually did [...] sometimes I'd mute or block them preemptively.*" –P9

Reducing receiver costs. Six participants (6/11) reported that they thought Sig was useful as a faster way to assess and filter accounts—they reported being able to check Sig's results within a few seconds. All participants mentioned the popup showing up to five tweets (Figure 4) was very useful as it quickly let them see the reason behind Sig's flagging, without having to scroll down to read tweets. Some participants also liked that Sig easily pulled up accounts' past replies, as the default Tweets tab does not include them (replies are included in Tweets & Replies tab). We believe this implies that Sig, and further S3s, can be used to reduce costs for accessing meaningful information from a person's history of posts.

"*...the extension would be a useful way for me to quickly get information without having to scroll back a zillion pages... "* –P11

"*Within a few seconds you could see what it was, and then you click it, it's a clear image of what's happening.*" –P10

Conventional social signals vs. S3s. At the same time, Sig seemed useful in gathering accurate information about accounts compared to conventional signals. Some participants (4/11) reported experiences when a verified account or a high-profile account with many followers was flagged as toxic or misinformation spreading by Sig. Participants said they tended to initially think of those accounts as "safe" or "reliable," but when seeing such accounts flagged by Sig, it made them think again about relying on conventional signals. For instance, P10 encountered a politician's account which was recommended by Twitter's algorithm. It had a blue verification badge but Sig flagged it as spreading misinformation. P10 said it made him realize that the politician "*had an agenda they were pushing*".

"*It was like I figured, cause she had a blue check mark and everything, I figured, and she's a politician so it's kind of funny when you see that [flagged by Sig as misinformation spreading].*" –P10

"*...if someone retweeted something and I looked on their profile and it was someone with a lot of followers, my instinctual reaction to that is like, 'Oh, they're probably fairly legit if they've got 30,000 followers,' but this made me rethink that, so those situations.*" –P5

Some participants even recalled their previous experience of noticing the discrepancy between an account's first impression and actual tweeting behavior days after following the account, and thought Sig would be useful to prevent such situations.

"*...I'll follow them and then weeks later or days later they'll say something and I'm like, 'Oh wow, okay. I shouldn't have followed this account,' and then I'll unfollow them and then maybe block them...* " –P7

Feeling safer. Three participants (3/11) reported that they felt safer when using Sig. When encountering toxic or misinformation-spreading accounts, these participants said they were glad they did not have to scroll further in the profile page and either left the page quickly or muted/blocked it. P9 said that they would keep the threshold low for "*general safety*" because they would rather risk seeing false positive accounts than miss legitimate flagging. Participants also said they can imagine Sig being especially useful for children, teenagers, and people who are often the target of online harassment and abuse [15]. Although a minority of participants reported feeling safer (3/11), we included this theme as online harassment can severely impact people, especially vulnerable populations [15].

"*I think people would feel it's safer, and so then they'll think twice before they post anything demeaning or bad, so it makes everything a little bit safer.*" –P3

"*I think I would keep setting it low for general safety, and then if something seemed like it was flagged incorrectly, then I could just note that to myself. I'd rather risk that than it not flagging people it should.*" –P9

Expanding S3s. All participants thought the two signals (toxicity and misinformation) were useful signals regarding social media's problem of online abusive behavior and misinformation. When asked if there were any signals other than toxicity and misinformation they wanted to add to the Sig, 9 participants suggested new social signals with 7 of them being S3s. Some participants speculated about fine-grained S3s for toxicity such as specific harassing behavior or triggering content posting behavior (e.g., content related to self-harm, sexually sensitive images). Other interviewees wished S3s for neutral signals (e.g., interests and hobbies), as it is hard to gauge a stranger account's interests "*when you're going through so many people on Twitter*" (P10).

"*I think this extension would be useful for anyone using Twitter. It's important to be able to see how much misinformation and toxicity you're allowing yourself to be exposed to and to have the ability to see that before it actual happens is powerful.*" –P4

"*... I'm wondering if it could also flag for other things, like maybe potentially triggering topics or something. I'm sure that would benefit a lot of people's experiences [...] I just really like the idea of being able to kind of see what an account posts about if you're interested in a flagging a certain subject. Maybe not even in a negative way, but just if you want to see more of a certain content on your timeline.*" – P1

S3s' downstream effect on communication. Donath has theorized that design changes that affect access to social signals may potentially alter communication dynamics [13]. When asked about their opinion on embedding S3s in the interface of social media platforms, participants had varying opinions. Some participants (4/11) were positive about the idea, especially about S3s for flagging misinformation. For instance, P9 envisioned Sig can help prevent misinformation from going viral. On the other hand some participants took a more measured view. They talked about issues including people finding ways to bypass S3s and algorithms never being perfect. Two participants that have heard about or know of machine learning were especially cautious about bias in machine learning algorithms. For example, P11 emphasized that machine learning can end up hurting vulnerable populations the most due to algorithmic bias, which echos prior work [33].

"*I think people might be a little more cautious talking to others and that it would be harder for misinformation to go viral if you could just check the account and it was all already flagged.*" –P9

"*... I wonder if people would find ways to transverse the extensions. I think it's just me being a pessimist in a sense of finding ways to state things in a way that don't flag toxic for example.*" –P2

Overall usability of Sig. Overall participants thought the extension's usability of was good in terms of design and speed. Participants were especially satisfied with Sig's visualization of S3s, rating it an average of 4.7/5 (Figure 6).

"*It was made for, like in cyber security we say, 'It's dumbed down for your grammar.' You know, anybody can understand it...*'" –P10

"*I thought it was a really good extension. It's something that if it was on the market as now, I'd have it already installed.* –P6"

DISCUSSION

We introduce the idea of synthesized social signals (S3s): social signals computationally derived from an account's history, and then rendered into the profile. Unlike conventional social signals, which are relatively easy to fake, S3s aim to reduce receiver cost and raise the cost of faking signals. To demonstrate the concept, we iteratively built Sig, an extensible Chrome extension that computes and visualizes S3s. Our field study showed that overall Sig reduced receiver cost and participants actively used Sig to mute, block, and decide whether or not to follow accounts. A few participants also reported they felt safe as a result. Next, we reflect on findings regarding the design of S3s and challenges of embedding S3s onto platforms.

Introducing New Social Signals in Social Platforms

Perhaps one of the most intriguing findings was that Sig identified accounts that are verified or high-profile (e.g. the politician's Twitter account that had a verification mark) as toxic or misinformation spreading. The participants that witnessed the flagged accounts felt that Sig rightly identified many of them (by looking at the modal that shows flagged tweets and further observing the profile page) and were surprised of the discrepancy between the conventional social signals (e.g. verification mark and high number of followers) and S3s. We believe the results imply that social signals like S3s can provide richer information to people while reducing costs. In short, S3s can be valuable in overcoming the dearth of social signals on social

platforms which cause people to experience discrepancy when bumping into accounts. The dearth may have to do with physical properties (compared to how f2f interactions), but also is deeply related to *how social platforms are designed to show primarily conventional social signals* (as shown in Figure 2). As Hollan and Stornetta [26] wrote in 1992, focusing on what online settings are differentially strong at may provide novel design insights. *How can we re-imagine and design social platforms in a way so the platforms fully utilize the strengths of online spaces?*

Empowering Users with Computation at the Edge

Our results show that participants liked that they could make their own decisions based on Sig. So far computation has been mostly hiding under the interface of platforms [19]. Our field study shows that surfacing results of computation as social signals enables people to make their own decisions, and that people are interested in such opportunities. For instance, participants wanted fine-grained S3s for toxicity, such as specific harassing behavior or triggering content posting behavior (e.g., sexually sensitive images) so they can easily avoid or even mute/block such accounts. Such empowerment can be important for platform users' safety, especially when platforms' current measures to address online harassment and abuse are insufficient [29, 40].

We also envision S3s can be useful for tackling misinformation by helping platform users themselves to easily discern fake accounts. Prior research has shown current social signals are not sufficient for aiding individuals to successfully identify false accounts from legitimate ones [48]. By augmenting S3s on top of profiles, tools like Sig can help people more easily identify fake accounts spreading misinformation. Furthermore, S3s fully utilize people's natural tendency of observing an account's posting behavior when gauging the account's credibility. For instance, participants perceived journalists that interact more frequently with Twitter followers (e.g., replying) as more credible [28].

Who Owns and Renders Social Signals?

Currently, Sig renders S3s in the browser, only for the user who has it installed. One question we can ask here is whether or not we want the platforms doing it instead. In other words, what if Twitter renders S3s instead of a browser extension? As Donath previously wrote, "One of the most intriguing possibilities of future social network technologies is that these social features can become part of of one's visible persona" [13], it is crucial to ask how S3s will affect people's communication with each other once existing on platforms. Participants held varying opinions about directly embedding S3s into social platforms. Some participants thought it would help people feel safer and said they did not mind other people inspecting their accounts. Others had mixed feelings mostly due to reasons like algorithms never being perfect, people finding ways to bypass algorithms, and self-censoring.

One possible solution to such different preferences is the one presented here: to build new tools for people that need S3s, that can augment existing platform interfaces. We argue that such tools can be a middle ground of letting users personalize their own social experiences by using S3s at the "edge," while at the same time preventing potential issues that can arise when embedding S3s entirely into a platform's interface. Moreover, we believe tools like Sig represent an interesting inversion of the usual role around data on social media platforms. Platform companies build and run algorithms using platform data for their own goals (e.g., serving better ads, increasing time-on-site, etc.), but they are invisible to users [19]. In contrast, Sig uses these data, which are originally typically archived and mined by companies, for human-centered goals—platform users can make social decisions based on the data in real-time interactions. We hope our work can spark new tools like Sig.

LIMITATIONS AND FUTURE WORK

We now turn to the potential negative impact of S3s and discuss future work that can help address such issues. A possible negative outcome is tools like Sig introducing algorithmic bias into the platforms [24, 33]. Furthermore, mislabeling accounts as being malicious could potentially escalate. Users might screenshot S3s and share them, which could lead to heated social conversations centering on the artifacts. One line of future work that can help address the issue of algorithmic bias is building tools like Sig in a way so that it is easy to gather users' feedback on the classification results. If tools like Sig provide features to let users easily flag and report accounts that they perceived as false positives or false negatives, such feedback can be used to improve the model or warn other users of the errors. Furthermore, it would be crucial to help users to easily set norms around how to make sense of and use S3s. Prior research have already suggested community-driven moderation tools to help groups of users set their own boundaries [5, 9]. Such features should be provided on the platforms to set effective norms around using S3s.

CONCLUSION

In order to address the problem of dearth of social signals on social media platforms, and the ease of faking them, we propose a new type of social signal called "synthesized social signal (S3s)." S3s are social signals derived from a person's history of posts using algorithms. Unlike conventional social signals, S3s aim to reduce receiver costs for deriving information using computation and increase the cost for faking information. To demonstrate the concept by using an iterative process, we built Sig, a Chrome extension that computes and visualizes S3s on social media platforms. The field study of Sig was conducted on Twitter with S3s for toxicity and misinformation-spreading behavior, two pressing problems on social platforms. Results show participants used Sig to make various judgements on stranger accounts. Participants reported they felt Sig provided useful and reliable information compared to conventional social signals.

ACKNOWLEDGMENTS

We thank Yoonjeong Cha, Woosuk Seo, John Joon Young Chung, Hari Subramonyam, Shagun Jhaver, and Joey Hsiao for their feedback. We also thank all participants and reviewers for their time. Im was supported by the National Science Foundation under grant IIP-1842949. Gilbert was supported by the National Science Foundation under grant IIS-1553376.

REFERENCES

[1] Marjolijn L Antheunis and Alexander P Schouten. 2011. The effects of other-generated and system-generated cues on adolescents' perceived attractiveness on social network sites. *Journal of Computer-Mediated Communication* 16, 3 (2011), 391–406.

[2] Saeideh Bakhshi, David A Shamma, and Eric Gilbert. 2014. Faces engage us: Photos with faces attract more likes and comments on instagram. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 965–974.

[3] Albert Bandura. 2001. Social cognitive theory: An agentic perspective. *Annual review of psychology* 52, 1 (2001), 1–26.

[4] Michael Barthel, Amy Mitchell, and Jesse Holcomb. 2016. Many Americans believe fake news is sowing confusion. *Pew Research Center* 15 (2016), 12.

[5] Lindsay Blackwell, Nicole Ellison, Natasha Elliott-Deflo, and Raz Schwartz. 2019. Harassment in Social Virtual Reality: Challenges for Platform Governance. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 100.

[6] Yazan Boshmaf, Ildar Muslukhov, Konstantin Beznosov, and Matei Ripeanu. 2011. The socialbot network: when bots socialize for fame and money. In *Proceedings of the 27th annual computer security applications conference*. ACM, 93–102.

[7] Alexandre Bovet and Hernán A Makse. 2019. Influence of fake news in Twitter during the 2016 US presidential election. *Nature communications* 10, 1 (2019), 7.

[8] Danah Michele Boyd. 2004. Friendster and publicly articulated social networking. In *Conference on Human Factors in Computing Systems: CHI'04 extended abstracts on Human factors in computing systems*, Vol. 24. 1279–1282.

[9] Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. 2018. The Internet's Hidden Rules: An Empirical Study of Reddit Norm Violations at Micro, Meso, and Macro Scales. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 32.

[10] Clayton Allen Davis, Onur Varol, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. 2016. Botornot: A system to evaluate social bots. In *Proceedings of the 25th International Conference Companion on World Wide Web*. International World Wide Web Conferences Steering Committee, 273–274.

[11] Marian Stamp Dawkins and Tim Guilford. 1991. The corruption of honest signalling. *Animal Behaviour* 41, 5 (1991), 865–873.

[12] Nicholas Diakopoulos, Munmun De Choudhury, and Mor Naaman. 2012. Finding and assessing social media information sources in the context of journalism. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 2451–2460.

[13] Judith Donath. 2007. Signals in social supernets. *Journal of Computer-Mediated Communication* 13, 1 (2007), 231–251.

[14] Judith Donath and Danah Boyd. 2004. Public displays of connection. *bt technology Journal* 22, 4 (2004), 71–82.

[15] Maeve Duggan. 2017. Online harassment 2017. (2017).

[16] Nicole Ellison, Rebecca Heino, and Jennifer Gibbs. 2006. Managing impressions online: Self-presentation processes in the online dating environment. *Journal of computer-mediated communication* 11, 2 (2006), 415–441.

[17] Nicole B Ellison, Jeffrey T Hancock, and Catalina L Toma. 2012. Profile as promise: A framework for conceptualizing veracity in online dating self-presentations. *new media & society* 14, 1 (2012), 45–62.

[18] Nicole B Ellison, Charles Steinfield, and Cliff Lampe. 2007. The benefits of Facebook "friends:" Social capital and college students' use of online social network sites. *Journal of computer-mediated communication* 12, 4 (2007), 1143–1168.

[19] Motahhare Eslami, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. 2015. I always assumed that I wasn't really that close to [her]: Reasoning about Invisible Algorithms in News Feeds. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. ACM, 153–162.

[20] Asle Fagerstrøm, Sanchit Pawar, Valdimar Sigurdsson, Gordon R Foxall, and Mirella Yani-de Soriano. 2017. That personal profile image might jeopardize your rental opportunity! On the relative impact of the seller's facial expressions upon buying behavior on Airbnb. *Computers in Human Behavior* 72 (2017), 123–131.

[21] Alan Grafen. 1990. Biological signals as handicaps. *Journal of theoretical biology* 144, 4 (1990), 517–546.

[22] Tim Guilford and Marian Stamp Dawkins. 1991. Receiver psychology and the evolution of animal signals. *Animal behaviour* 42, 1 (1991), 1–14.

[23] Tim Guilford and Marian Stamp Dawkins. 1995. What are conventional signals? *Animal Behaviour* 49, 6 (1995), 1689–1695.

[24] Sara Hajian, Francesco Bonchi, and Carlos Castillo. 2016. Algorithmic bias: From discrimination discovery to fairness-aware data mining. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 2125–2126.

[25] Jeffrey T Hancock and Philip J Dunham. 2001. Impression formation in computer-mediated communication revisited: An analysis of the breadth and intensity of impressions. *Communication research* 28, 3 (2001), 325–347.

[26] Jim Hollan and Scott Stornetta. 1992. Beyond being there. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 119–125.

[27] Benjamin D Horne, Sara Khedr, and Sibel Adali. 2018. Sampling the news producers: A large news and feature data set for the study of the complex media landscape. In *Twelfth International AAAI Conference on Web and Social Media*.

[28] Mi Rosie Jahng and Jeremy Littau. 2016. Interacting is believing: Interactivity, social cue, and perceptions of journalistic credibility on twitter. *Journalism & Mass Communication Quarterly* 93, 1 (2016), 38–58.

[29] Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. 2018. Online harassment and content moderation: The case of blocklists. *ACM Transactions on Computer-Human Interaction (TOCHI)* 25, 2 (2018), 12.

[30] Cliff Lampe, Nicole Ellison, and Charles Steinfield. 2007. A familiar face (book): profile elements as signals in an online social network. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 435–444.

[31] Kevin Munger. 2017. Tweetment effects on the tweeted: Experimentally reducing racist harassment. *Political Behavior* 39, 3 (2017), 629–649.

[32] Dietmar Offenhuber and Judith Donath. 2007. Comment flow: visualizing communication along network path. *Poster presented at IEEE InfoVis* 7 (2007).

[33] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 1668–1678.

[34] Cuihua Shen, Peter Monge, and Dmitri Williams. 2014. Virtual brokerage and closure: Network structure and social capital in a massively multiplayer online game. *Communication Research* 41, 4 (2014), 459–480.

[35] Aaron Smith, Laura Silver, Shawnee Cohn, and Stefan Cornibert. 2019. Publics in Emerging Economies Worry Social Media Sow Division, Even as They Offer New Chances for Political Engagement. (2019).

[36] John Maynard Smith and DGC Harper. 1988. The evolution of aggression: can selection generate variability? *Philosophical Transactions of the Royal Society of London. B, Biological Sciences* 319, 1196 (1988), 557–570.

[37] John Maynard Smith and David Harper. 2003. *Animal signals*. Oxford University Press.

[38] Maynard J Smith and David GC Harper. 1995. Animal signals: models and terminology. *Journal of theoretical biology* 177, 3 (1995), 305–311.

[39] Michael Spence. 1978. Job market signaling. In *Uncertainty in economics*. Elsevier, 281–306.

[40] Nitasha Tiku and Casey Newton. 2015. Twitter CEO: "We suck at dealing with abuse.". *The Verge* 4, 2 (2015), 2015.

[41] Stephanie Tom Tong, Brandon Van Der Heide, Lindsey Langwell, and Joseph B Walther. 2008. Too much of a good thing? The relationship between number of friends and interpersonal impressions on Facebook. *Journal of computer-mediated communication* 13, 3 (2008), 531–549.

[42] Sonja Utz. 2010. Show me your friends and I will tell you what type of person you are: How one's profile, number of friends, and type of friends influence impression formation on social network sites. *Journal of Computer-Mediated Communication* 15, 2 (2010), 314–335.

[43] Onur Varol, Emilio Ferrara, Clayton A Davis, Filippo Menczer, and Alessandro Flammini. 2017. Online human-bot interactions: Detection, estimation, and characterization. *arXiv preprint arXiv:1703.03107* (2017).

[44] Thorstein Veblen. 2017. *The theory of the leisure class*. Routledge.

[45] Joseph B Walther. 1996. Computer-mediated communication: Impersonal, interpersonal, and hyperpersonal interaction. *Communication research* 23, 1 (1996), 3–43.

[46] Joseph B Walther and Malcolm R Parks. 2002. Cues filtered out, cues filtered in. *Handbook of interpersonal communication* 3 (2002), 529–563.

[47] Joseph B Walther, Brandon Van Der Heide, Lauren M Hamel, and Hillary C Shulman. 2009. Self-generated versus other-generated statements and impressions in computer-mediated communication: A test of warranting theory using Facebook. *Communication research* 36, 2 (2009), 229–253.

[48] Gang Wang, Manish Mohanlal, Christo Wilson, Xiao Wang, Miriam Metzger, Haitao Zheng, and Ben Y Zhao. 2012. Social turing tests: Crowdsourcing sybil detection. *arXiv preprint arXiv:1205.3856* (2012).

[49] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 1391–1399.

[50] Amotz Zahavi. 1977. The cost of honesty (further remarks on the handicap principle). *Journal of theoretical Biology* 67, 3 (1977), 603–605.

[51] Aaron Robert Zinman. 2011. *Me, myself, and my hyperego: understanding people through the aggregation of their digital footprints*. Ph.D. Dissertation. Massachusetts Institute of Technology.