Managing Social (Media) Signals: Using Platform Labelers to Improve AI- and Social Media-Mediated Communication

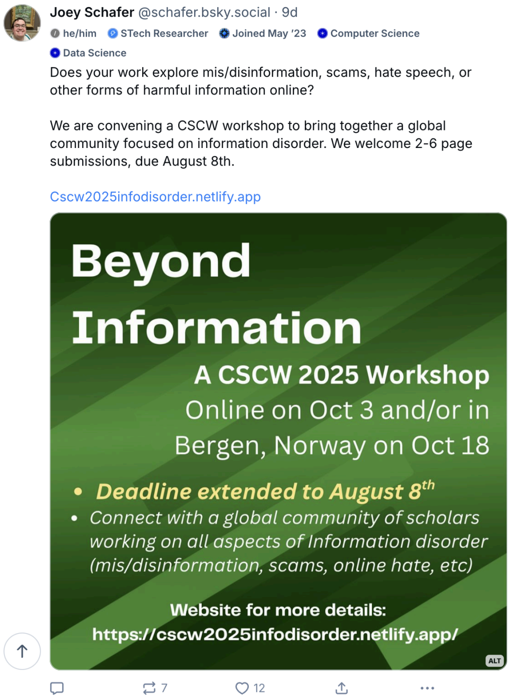
Joseph S. Schafer
SETS Fellow, Cornell Tech
PhD Candidate, University of Washington, Seattle

Motivation

Social media systems are incredibly important to daily life, including for political discussions, disaster relief, and community-building. However, especially in the age of large language models which can plausibly sound human online, differentiating who one is engaging with online is a critical challenge to protect users from information operations, scams, and harassment, among other possible harms.

One approach which could help inform tool-building to facilitate these differentiations derives from signaling theory, where people use signals to understand the capabilities, identities, and behaviors of those whom they engage. For this fellowship, I explored how signaling theory has been used in previous social media research, and built a prototype system for creating automated, synthesized signals on the Bluesky platform to aid user understandings of other accounts on the network.

Bluesky is a decentralized microblogging platform with about 38 million accounts. The platform has gotten attention (and a large portion of its users) due to challenges over moderation policies, trust and safety, and legal access that users had with other microblogging platforms, like Twitter/X . As a result, this community may be particularly interested in T&S interventions. Additionally, Bluesky has a composable, decentralized moderation system, allowing third parties to create natively-rendered labels. An example of how labelers are rendered on the platform is shown in *Figure 1.*



Joey Schafer @schafer.bsky.social · 9d

Does your work explore mis/disinformation, scams, hate speech, or other forms of harmful information online?

We are convening a CSCW workshop to bring together a global community focused on information disorder. We welcome 2-6 page submissions, due August 8th.

Cscw2025infodisorder.netlify.app

Beyond Information

A CSCW 2025 Workshop

Online on Oct 3 and/or in Bergen, Norway on Oct 18

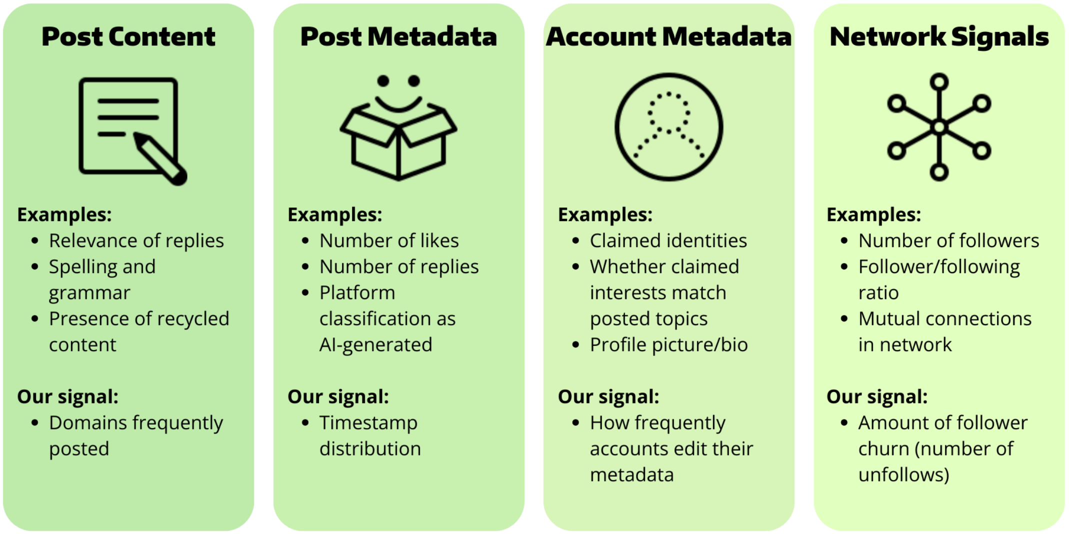
- *Deadline extended to August 8th*
- *Connect with a global community of scholars working on all aspects of Information disorder (mis/disinformation, scams, online hate, etc)*

Website for more details: https://cscw2025infodisorder.netlify.app/

Figure 1: A screenshot of labels applied to a post of the author's.

Kinds of Signals

I first conducted a review of several papers on social signals, bot and inauthentic behavior detection, and credibility indicators used by social media users. Broadly, the signals fit into four categories: post content, post metadata, account metadata, and network signals, which we outline examples of (and our focused, synthesized signals) below.



Post Content

Examples:
- Relevance of replies
- Spelling and grammar
- Presence of recycled content

Our signal:
- Domains frequently posted

Post Metadata

Examples:
- Number of likes
- Number of replies
- Platform classification as AI-generated

Our signal:
- Timestamp distribution

Account Metadata

Examples:
- Claimed identities
- Whether claimed interests match posted topics
- Profile picture/bio

Our signal:
- How frequently accounts edit their metadata

Network Signals

Examples:
- Number of followers
- Follower/following ratio
- Mutual connections in network

Our signal:
- Amount of follower churn (number of unfollows)

Icons from icons8

Figure 2: A chart showing the four categories of signals we identified, examples in each category, and the synthesized signals we are initially implementing.

One struggle when considering how to use automated signals is how to make these indicators more quantitative, as many of the signals identified in past literature are either inferred via ML models, or are more qualitative senses, such as how subjective social media content a user posts is. For our implementations, we are focusing on signals which are highly transparent and objective, at the expense of making them more value-laden or evaluative. We are also not attempting to have labels which directly identify accounts, e.g. “this is a bot” or “this is a scam” labels. Instead, we focus on labeling specific, observable behaviors, and allowing these behaviors to serve as signals for users to determine whether they would want to engage with accounts with those behaviors.

A key reason for this decision is that some of these behaviors, even if highly unusual, do not *necessarily* mean that an account is suspicious or nefarious. For example, when looking at frequency of bio changes as a noteworthy signal, we found accounts which have extremely frequent updates because they auto-update their bio to include, among other things, the time of day, their most recent follower, the temperature at the account holder’s location, and currency exchange rates. As a result, the signals we are applying are not final value-laden judgments about the account, but instead are just additional information and context for users.

Bluesky Labeler Architecture

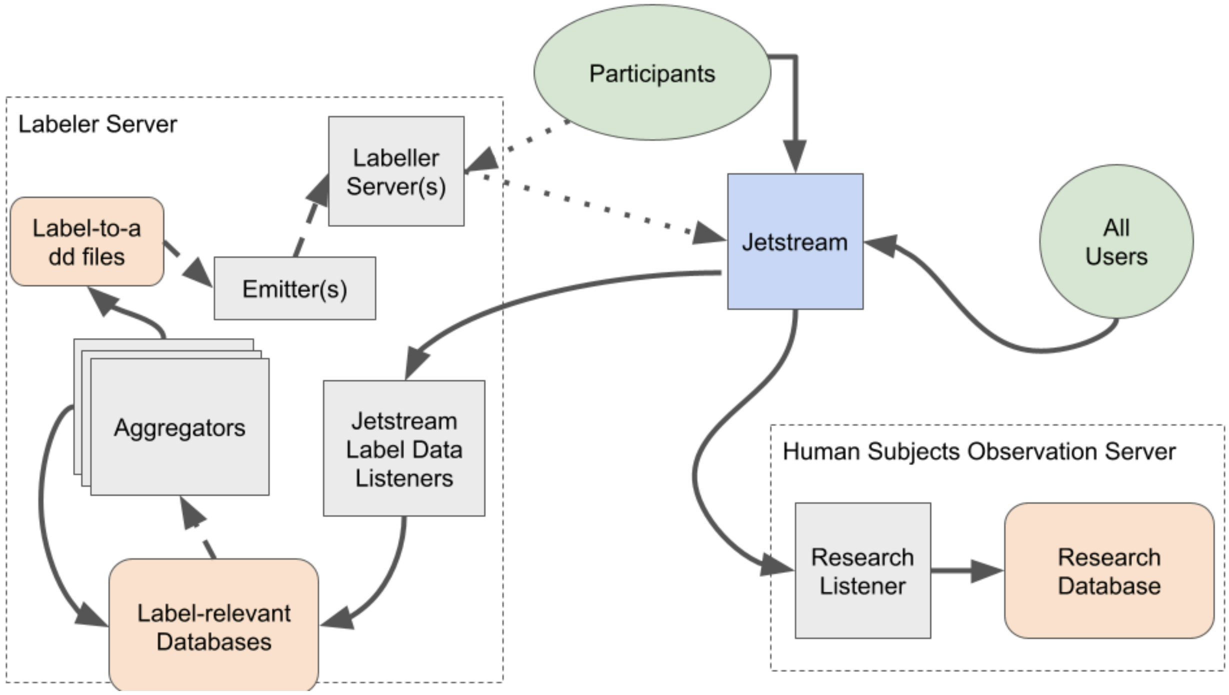


Figure 3: Architecture of our synthesized signal-based labeling system

To enable these automated signals, I built a prototype system which uses data from the Bluesky jetstream (a firehose of updates to the network including all new information posted) and synthesizes this information into labels which can be viewed by users of the platform. Figure 3, above, illustrates this total architecture.

In brief, we set up several *listeners*, which are Node.js scripts which track particular kinds of events on the firehose (such as new follows and unfollows, or profile metadata updates), and save these updates to a SQLite database. This database is then queried and processed daily by the *aggregators*, which determine which accounts have frequently engaged in behavior which should be labeled (e.g. editing their bio too frequently, or unfollowing too many accounts), and saves these as text files. These are then read by the *emitter*, which is another Node.js script which tells the *label server* (itself another Node-based script, based off of this repository) to make these labels publicly visible on the Bluesky network. Currently, all of these are running off of a Google Cloud Platform server.

This architecture is designed to allow for efficient scaling, since new listeners and aggregators can easily be added to the pipeline to account for new signals, and emitters can be structured to send to multiple label servers, or selectively send to separate servers, for multiple experimental arms. Using this firehose also allows for tracking the behavior of users who subscribe to the labeler, which we plan to use for our future human subjects research on this topic described below.

Next Steps

Human Subjects Testing

Beyond building these technical systems for increasing the visibility of behavioral signals by accounts, it is critical to understand how users actually view these signals, how it changes their beliefs about the Bluesky platform, and how it impacts their behavior. To build on this work, we plan to run some field experiments, where we have participants use ("subscribe") to one of several labelers (one per experimental arm), and see if this impacts their behaviors on the platform. In particular, we will be measuring 1) whether having these labels be active changes users' overall frequency of engagement, and 2) whether it increases or decreases their engagement with labelled users (such as blocking, unfollowing, liking/reposting/commenting less, etc.).

To understand how these impact users' affective experiences beyond quantitative behavioral measures, we will do follow-up surveys and interviews with participants to understand how they make sense of these labels, the accounts who are labeled, and the overall Bluesky ecosystem with these additional signals of transparency baked in.

Building Additional Signals and Fine-Tuning Thresholds.

Alongside our human subjects testing, we are attempting to build out additional automated signals of behavior beyond the four described above, such as how many labeled accounts an account follows, how often accounts delete posts, and if post content includes filler text that would indicate it is AI-generated (such as starting a post with "As a language model…"). Additionally, we are still exploring how best to calibrate the exact thresholds for our signals, such as the amount of follower churn or account metadata edits required to be labeled.