RESEARCH STATEMENT **SIQI WU**

We are living in an era where people's perception and behavior are increasingly influenced by the digital world. Public discourse on controversial topics may evoke emotional harm or even incite physical violence to some groups; the spread of misinformation may change election results; and extreme content on social media may polarize or radicalize people. To this end, new understanding of contemporary social phenomena is urgently needed.

I am a researcher in **computational social science (CSS)** who collects, models, and analyzes social data at scale. In my doctoral and postdoctoral work, I produced descriptive measurement results on topics such as **content consumption** and **political polarization**, and contributed a variety of **resources** that aid in these measurements (Figure 1). Specifically, I conceptualized a framework to characterize online consumption behaviors, and I identified distinct behavioral patterns between the political left and right.

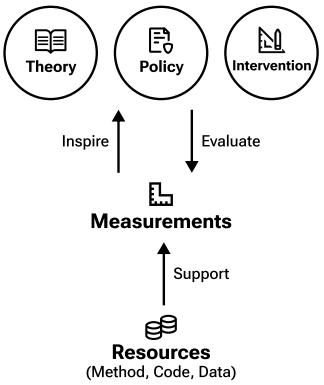

Figure 1. My contributions to CSS research through conducting large-scale measurements and developing open resources.

Large-scale measurements are valuable on two fronts. First, they have the potential to transform our understanding of large-scale social phenomena, which can inspire the creation of theories and the design of policies and interventions. Second, they also provide empirical evidence for evaluating theories, policies, and interventions. The main thrust that enables such measurements is the continuous development of new methods, models, and open datasets. While my primary focus is on revealing patterns of digital traces to advance our knowledge about the human society, I also envision more broadly the next generation of social systems that are safe, fair, responsible, and transparent. To achieve this goal, I will pursue collaborations with computer scientists, HCI researchers, sociologists, policy scholars, as well as private sectors, to use my expertise in large-scale measurements to empower the creation and evaluation of data-driven theories, policies, and interventions.

1 MEASUREMENT CONTRIBUTIONS

1.1 A two-step framework to model content consumption

Human attention is a scarce resource. In online platforms, while users have an unprecedented volume of content to choose from, the content competes for limited user attention. I conceptualize online consumption actions as two steps: the first is based on appeal, measured by clicks or views; the second is based on quality, measured by post-click metrics, e.g., dwell time, likes, or shares. I have applied this conceptual framework to measure the collective video consumption patterns on YouTube and explore the predictability of each metric. These measurements can inform content creators of engaging topics, inspire designs for filtering out low-quality content, and shed light on prioritizing quality content in recommender systems.

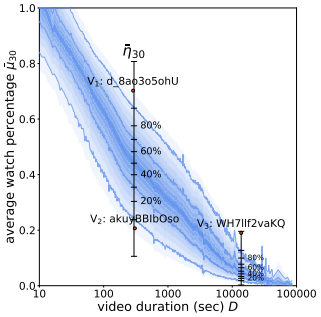

Figure 2. Engagement map and relative engagement.

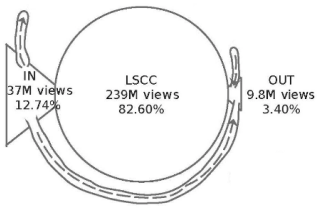

Figure 3. Structure of YouTube video recommendation networks. Attention can only flow along a single direction, from IN to LSCC to OUT.

How to measure the quality of videos? [11] I studied the time and percentage of videos being watched on YouTube. Video length is an important covariate for the watching metrics because longer videos generally make the users stay for a longer time but are less likely to keep them watching till the end. To calibrate this effect, I designed a two-dimensional tool, called *engagement map* (Figure 2), which captures the nonlinear relationship between video length and watch percentage. Based on it, I proposed a new metric, called *relative engagement*, as the watch percentage rank percentile among videos of similar lengths. I show that this metric is closely correlated with recognized notions of quality, stable over time, and predictable even before videos' upload. While generic topics (e.g., news, music) are not predictive of the relative engagement metric; specific topics (e.g., Obama, Indian musical) are somewhat predictive of it.

How do the video recommendation networks look like? [7, 12] I studied the structure of video networks built by the YouTube recommender system, and connected it to the dynamics of video popularity. I find that when viewing a video, users are more likely recommended to videos more viewed than the original. This mechanism is one factor that explains the well-known "rich get richer" phenomenon, whereby popular items tend to attract even more views. I also characterized the recommendation networks by classifying them into three components: links point from the IN component to the Largest Strongly Connected Component (LSCC) to OUT (Figure 3). The core LSCC consists of only 23% of videos but occupies 83% of views. The structure outlines pathways for the human attention (metricized by views) to flow in. Finally, I find that network structure is a useful feature in predicting video view counts.

Together with a collaboration work that models the impact of Twitter shares on video views [2], this two-step framework is particularly useful for explaining the predictability of online consumption metrics: the pre-click metrics are impacted by the position in the recommendation networks and external promotion, making them less predictable; while the post-click metrics are only relevant to the intrinsic content quality, making them highly predictable.

1.2 Ideological asymmetries between political left and right

Political polarization is an important area where scholars from multiple disciplines have made efforts to understand the psychological diversity of, design safe conversation space for, and bridge the information gap between partisan groups. In my research, I rely on observational data to study user behavior and platform effects, with a focus on the differences between the political left and right. Specifically, how they talk to each other, read news, and spread information. My work reveals many interesting patterns of group-level behavioral discrepancy.

How do partisan users engage in online discourse? [10] I collected a large dataset of YouTube political videos from US partisan media and millions of comments on them to investigate cross-partisan discussions. Contrary to a simple narrative of selective exposure, I find a surprising amount of cross-talk: most users posted at least once on both left-leaning and right-leaning YouTube channels. Cross-talk, however, is not symmetric. Conservatives are much more likely to comment on left-leaning videos than liberals on right-leaning videos (Figure 4). Secondly, YouTube's comment sorting algorithm makes cross-partisan comments modestly less visible. Lastly, using Perspective API's toxicity scores, I find evidence against the hypothesis that the asymmetry merely reflects conservatives trolling on liberal space. I find that conservatives' comments are not significantly more toxic than those of liberals overall. However, when users reply to other users, cross-partisan replies are indeed more toxic than co-partisan replies, on both left-leaning and right-leaning videos, with cross-partisan replies being especially toxic when the video political leaning aligns with the replier's—a phenomenon that we coin the term "defense of home territory".

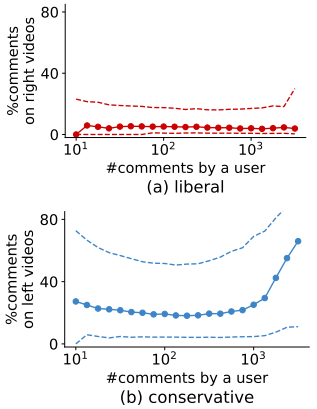

Figure 4. Cross-partisan comments made by liberals and conservatives. Solid (dashed) line indicates median (inter-quartile range).

How do partisan users consume news? [14] I led a project that profiles the audience's political leaning distribution for news media in eight countries. I find that while left-leaning media have a non-trivial amount of right-leaning audiences; right-leaning media have few left-leaning audiences. This resonates with my previous observation that conservatives comment more on cross-partisan content. I also find that the more extreme the media is, the fewer cross-partisan audience it would attract. This work contributes a new set of media bias scores by averaging the political leaning scores of users who have shared URLs from the media. These new scores not only correlate with existing media bias reporting for sources that have been assessed previously, but also provide estimates for unevaluated sources. This resource offers enormous research opportunities for scholars who want to study media consumption beyond the US (Figure 5).

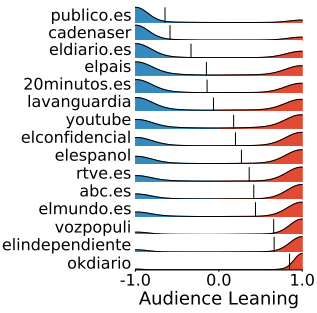

Figure 5. Audience leaning distribution for the top 15 media in Spain. Only two media have bias ratings from MBFC or AllSides.

In addition, with collaborators and mentees, we used several video-centric metrics to characterize how online attention is accumulated for the two ideological groups across platforms, across topics, and over time [3]. We also designed an audit study to measure the empirical effects of platform buttons on removing unwanted recommendations on YouTube [4].

2 RESOURCE CONTRIBUTIONS

Open datasets. I have released three YouTube datasets [10–12] and three Twitter datasets [5, 13, 14]. These datasets contain millions of YouTube videos and billions of tweets. My YouTube datasets are unique in a way that they include the daily time series of video view count and watch time. My Twitter datasets were specifically

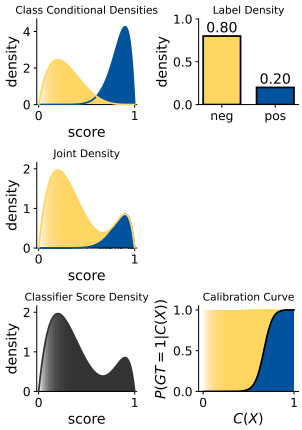

Figure 6. Joint distributions generated by *pyquantifier*.

constructed to minimize the Twitter sampling effects. Those datasets have been used by the research community to study topics such as remote learning, information diffusion, and stance detection.

Methods and demonstrations. With collaborators, we designed a deep learning model for cyberbullying detection [1]. We explored ways to improve the training process in crowdsourced labeling [8]. We also published two demos: HIPie is an interactive system that predicts the video view count given an input of tweet promotion [2]; AttentionFlow is a tool that visualizes the influence in networks of time series data [7].

A package for prevalence estimation. I developed a Python package called *pyquantifier*. It provides built-in functions to model joint distributions of a labeled dataset and outlines the distribution stability assumptions used for extrapolation to future datasets. This package was first presented at ICWSM 2023 as a tutorial [9]. It is also used in a Center for Social Media Responsibility project that monitors the prevalence of unhealthy conversations on social media [6].

3 FUTURE RESEARCH

My research goal is to advance our understanding of contemporary social phenomena via large-scale measurements and design future social systems via policies and interventions driven by data.

Structure, motivation, and reasoning of cross-talk. My previous work [10] contributes a cross-partisan discussion dataset on YouTube. I want to analyze the conversation—human to human messaging—in this dataset by characterizing the conversation structure, parsing the texts via computational methods, and understanding the user motivation via in-depth interviews. Are conservatives posting on the left-leaning channels to inform, convince, entertain, or learn? Do people still participate in cross-partisan conversations if their comments receive no engagement? Why do people use more toxic languages when talking on the ideological aligned space? Those questions are crucial to understanding cross-cutting communication and reducing conflicts in this politically segregated society.

News production and consumption in news deserts. My previous work [14] investigates news consumption in eight countries. I want to expand the research to areas where news is not locally available, dubbed news deserts. In the US, rural areas are becoming news deserts as local outlets shut down. Internationally, entire countries can become news deserts due to underdeveloped news industry. One alarming example is in the early days of COVID-19 pandemic, the whole world relied on medical information and advice from a few developed countries. What media supply news to the news deserts? What news are local audiences demanding? These questions are important for understanding cultural and social influence, and can inform the design of interventions that irrigate the deserts.

REFERENCES

[1] Lu Cheng, Kai Shu, **Siqi Wu**, Yasin N Silva, Deborah L Hall, and Huan Liu. 2020. Unsupervised Cyberbullying Detection via Time-Informed Gaussian Mixture Model. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*.

[2] Quyu Kong, Marian-Andrei Rizoiu, **Siqi Wu**, and Lexing Xie. 2018. Will This Video Go Viral? Explaining and Predicting the Popularity of YouTube Videos. In *Companion Proceedings of the The Web Conference*.

[3] JooYoung Lee*, **Siqi Wu***, Ali Mert Ertugrul*, Yu-Ru Lin, and Lexing Xie. 2022. Whose Advantage? Measuring Attention Dynamics across YouTube and Twitter on Controversial Topics. In *Proceedings of the 16th International AAAI Conference on Web and Social Media*.

[4] Alexander Liu, **Siqi Wu**, and Paul Resnick. 2024. How to Train Your YouTube Recommender to Avoid Unwanted Videos. In *Proceedings of the 18th International AAAI Conference on Web and Social Media*.

[5] Juergen Pfeffer, Daniel Matter, Kokil Jaidka, Onur Varol, Afra Mashhadi, Jana Lasser, Dennis Assenmacher, **Siqi Wu**, Diyi Yang, Cornelia Brantner, et al. 2023. Just Another Day on Twitter: A Complete 24 Hours of Twitter Data. In *Proceedings of the 17th International AAAI Conference on Web and Social Media*.

[6] Paul Resnick, **Siqi Wu**, and Vitaliy Lyapota. [n. d.]. The H|O|T News Comments Metric. *Report for the Center for Social Media Responsibility* ([n. d.]).

[7] Minjeong Shin*, Alasdair Tran*, **Siqi Wu***, Alexander Mathews, Rong Wang, Georgiana Lyall, and Lexing Xie. 2021. AttentionFlow: Visualising Influence in Networks of Time Series. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*.

[8] **Siqi Wu**, Ashwin Rajadesingan, Ceren Budak, Kelly Garrett, Daniel Sude, and Paul Resnick. [n. d.]. ACT: Active Crowd Training for Crowdsourced Labeling. *Working paper* ([n. d.]).

[9] **Siqi Wu** and Paul Resnick. [n. d.]. Calibrate-Extrapolate: Rethinking Prevalence Estimation with Black Box Classifiers. *Working paper* ([n. d.]).

[10] **Siqi Wu** and Paul Resnick. 2021. Cross-Partisan Discussions on YouTube: Conservatives Talk to Liberals but Liberals Don't Talk to Conservatives. In *Proceedings of the 15th International AAAI Conference on Web and Social Media*. **Spotlight Paper (top 8)**.

[11] **Siqi Wu**, Marian-Andrei Rizoiu, and Lexing Xie. 2018. Beyond Views: Measuring and Predicting Engagement in Online Videos. In *Proceedings of the 12th International AAAI Conference on Web and Social Media*.

[12] **Siqi Wu**, Marian-Andrei Rizoiu, and Lexing Xie. 2019. Estimating Attention Flow in Online Video Networks. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019). **Best Paper Honorable Mention (top 5%)**.

[13] **Siqi Wu**, Marian-Andrei Rizoiu, and Lexing Xie. 2020. Variation across Scales: Measurement Fidelity under Twitter Data Sampling. In *Proceedings of the 14th International AAAI Conference on Web and Social Media*.

[14] Cai Yang, Lexing Xie, and **Siqi Wu**^. 2023. The Shapes of the Fourth Estate During the Pandemic: Profiling COVID-19 News Consumption in Eight Countries. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023).