Siqi Wu

Ann Arbor, MI, USA | siqiwu@umich.edu | https://avalanchesiqi.github.io/

RESEARCH SUMMARY

I aim to derive principles for responsible social systems, with my core expertise in large-scale quantitative analysis and collaboration with experts in social science, communication, and political science. I work towards my goal through a variety of methods – from curating unique social media datasets, to modeling collective user behavior, to building algorithmic audits for identifying negative platform effects, to designing and testing new HCI systems.

PROFESSIONAL EXPERIENCE

Research Consultant | University of Michigan.^ *2020.09 – present*

^ Responsibilities equivalent to a research fellow, work conducted remotely due to pandemic.

- Center for Social Media Responsibility. Adviser: Paul Resnick

Projects include (1) measuring cross-partisan discussions on YouTube; (2) characterizing concept difference and (3) quantifying the impacts of annotator demographics in labeling hate, offensive, and toxic social media content; (4) calibrating prevalence estimation methods for online problematic speech; (5) designing effective training procedure for crowdsourced labeling; (6) auditing polarized recommendations in YouTube recommender system.

Research Fellow | Australian National University. Canberra, ACT, Australia *2020.09 – 2021.03*

- Computational Media Lab. Adviser: Lexing Xie

Projects include (1) measuring collective human attention across different social platforms; (2) picturing global media consumption during COVID-19.

Software Engineer | MicroStrategy, Inc. Hangzhou, Zhejiang, China *2015.09 – 2016.05*

- Big Data Engine team. Built Apache Spark alike engine to process massive data.

Software Developer Intern | Baidu, Inc. Beijing, China *2014.12 – 2015.02*

- Baidu Maps team. Developed a tool to collect realtime traffic status.

EDUCATION

Australian National University. Canberra, ACT, Australia *2016.06 – 2020.09*

- PhD in Computer Science. Advisers: Lexing Xie and Marian-Andrei Rizoiu
- Thesis: Measuring collective attention in online content: Sampling, engagement, and network effects

University of Melbourne. Melbourne, VIC, Australia *2013.07 – 2015.07*

- Master of Information Technology. Adviser: Richard Sinnott
- Thesis: An architecture for big data processing and visualisation of traffic data

Tianjin University. Tianjin, China *2008.09 – 2012.06*

- Bachelor of Electronics Engineering. Adviser: Yugong Wu

AWARDS

- AAAI ICWSM Best Paper Finalist *'21*
- ACM CSCW Best Paper Honorable Mention *'19*
- Google PhD Fellowship *'18*

PUBLICATIONS

* Equal authorship

14. YouTube users unaware of effective ways to remove unwanted recommendations

Liu, **Wu**, and Resnick. *Under review*

Abstract: YouTube provides features for users to indicate disinterest when presented with unwanted recommendations, such as the "Not interested" and "Don't recommend channel" buttons.These buttons are purported to allow the user to correct "mistakes" made by the recommendation system. Yet, relatively little is known about the empirical efficacy of these buttons. Neither is much known about users' awareness of and confidence in them. To address these gaps, we simulated YouTube users with sock puppet agents. Each agent first executed a "stain phase", where it watched many videos of one assigned topic; then it executed a "scrub phase", where it tried to remove recommendations of the assigned topic. Each agent repeatedly applied a single scrubbing strategy, either indicating disinterest in one of the videos visited in the stain phase (disliking it or deleting it from the watch history), or indicating disinterest in a video recommended on the homepage (clicking the "not interested" or "don't recommend channel" button or opening the video and clicking the dislike button). We found that the stain phase significantly increased the fraction of the recommended videos on the user's homepage dedicated to the assigned topic. For the scrub phase, using the "Not interested" button worked best, significantly reducing such recommendations in all topics tested, on average removing 88% of them. Neither the stain phase nor the scrub phase, however, had much effect on videopage recommendations (those given to users while they watch a video). We also ran a survey (N =300) asking adult YouTube users in the US whether they were aware of and used these buttons before, as well as how effective they found these buttons to be. We found that 44% of participants were not aware that the "Not interested" button existed. However, those who were aware of this button often used it to remove unwanted recommendations (82.8%) and found it to be modestly effective (3.42 out of 5).

13. The shapes of the fourth estate during the pandemic: Profiling COVID-19 news consumption in eight countries

Yang, Xie, and **Wu**. *Under review*

Abstract: News media is often referred to as the Fourth Estate, a recognition of their political power. New understandings of how media shape political beliefs and influence collective behaviors are urgently needed in an era when public opinion polls do not necessarily reflect election results and users influence each other in real-time under algorithm-mediated content personalization. In this work, we measure not only the average but also the distribution of audience political leanings for different media across different countries. The methodological components of these new measurements include a high-fidelity COVID-19 tweet dataset; high-precision user geolocation extraction; and user political leaning estimated from the within-country retweet networks involving politicians. We focus on geolocated users from eight countries, profile user leaning distribution for each country, and analyze bridging users who have interactions between multiple countries. Except for France and Turkey, we observe consistent bi-modal user leaning distributions in the other six countries, and find that cross-country retweeting behaviors do not oscillate across the partisan divide. More importantly, this study contributes a new set of media bias estimates by averaging the leaning scores of users who share URLs from this media domain. Through two validations, we find that the new average audience leaning scores strongly correlate with existing media bias scores. Lastly, we profile the COVID-19 news consumption by examining the audience distribution for top media in each country, and for selected media across all countries. Those analyses help answer questions such as: Does center media reuters have a more balanced audience base than partisan media cnn and foxnews in the US? Does far-right media breitbart attract any left-leaning readers in any countries? Does cnn reach a more balanced audience base in the US than in UK and Spain? In sum, our data-driven methods allow us

to study media that are not often collected in editor-curated media bias reporting, especially in non-English-speaking countries. We hope that such cross-country research would inform media outlets of their effectiveness and audience bases in different countries, inform non-government and research organizations about the country-specific media audience profiles, and inform individuals to reflect on our day-to-day media diet.

12. **Prevalence estimation in social media using black box classifiers**
    Wu and Resnick. *ICWSM '23*. Tutorial
    Abstract: Many problems in computational social science require estimating the proportion of items with a particular property. This counting task is called prevalence estimation or quantification. Frequently, researchers have a pre-trained classifier available to them. However, it is usually not safe to simply apply the classifier to all items and count the predictions of each class, because the test dataset may differ in important ways from the dataset on which the classifier was trained, a phenomenon called distribution shift. In addition, a second type of distribution shift may occur when one wishes to compare the prevalence between multiple datasets, such as tracking changes over time. To cope with that, some assumptions need to be made about the nature of possible distribution shifts across datasets.

    This tutorial will introduce an end-to-end framework for prevalence estimation using black box (pre-trained) classifiers, with a focus on applications in social media. The framework consists of a calibration phase and an extrapolation phase, aiming to address the two types of distribution shifts described above. We will provide hands-on exercises that walk the participants through solving a real world problem of quantifying positive tweets in datasets from two separate time periods. All datasets, pre-trained models, and example codes will be provided in a Jupyter notebook. After attending this tutorial, participants will be able to understand the basics of the prevalence estimation problem in social media, and construct a data analysis pipeline to conduct prevalence estimation for their own projects.

11. **Just another day on Twitter: A complete 24 hours of Twitter data**
    Pfeffer, Matter, Jaidka, Varol, Mashhadi, Lasser, Assenmacher, **Wu**, Yang, Brantner, Romero, Otterbacher, Schwemmer, Joseph, Garcia, and Morstatter. *ICWSM '23*. Dataset paper
    Abstract: At the end of October 2022, Elon Musk concluded his acquisition of Twitter. In the weeks and months before that, several questions were publicly discussed that were not only of interest to the platform's future buyers, but also of high relevance to the Computational Social Science research community. For example, how many active users does the platform have? What percentage of accounts on the site are bots? And, what are the dominating topics and sub-topical spheres on the platform? In a globally coordinated effort of 80 scholars to shed light on these questions, and to offer a dataset that will equip other researchers to do the same, we have collected all 375 million tweets published within a 24-hour time period starting on September 21, 2022. To the best of our knowledge, this is the first complete 24-hour Twitter dataset that is available for the research community. With it, the present work aims to accomplish two goals. First, we seek to answer the aforementioned questions and provide descriptive metrics about Twitter that can serve as references for other researchers. Second, we create a baseline dataset for future research that can be used to study the potential impact of the platform's ownership change.

10. **Whose advantage? Measuring attention dynamics across YouTube and Twitter on controversial topics**
    Lee*, **Wu***, Ertugrul*, Lin, and Xie. *ICWSM '22*. Full paper
    Abstract: The ideological asymmetries in contested online spaces have been recently observed, where conservative voices seem to be relatively more pronounced even though liberals have the population advantage on digital platforms. Most of the prior works, however, focused on either a single platform or a

single political topic. Whether one ideological group garners more attention across platforms and how the attention dynamics evolve have not been explored. In this work, we present a quantitative study that links collective attention across two social media platforms -- YouTube and Twitter, centered on online activities surrounding videos of controversial political topics including Abortion, Gun control, and Black Lives Matter over 16 months. We propose a set of video-centric metrics to characterize how online attention is accumulated from different political groups over time. Contrasting with prior observations, we find that neither ideological side is on a winning streak: left-leaning videos are overall more viewed, more engaging, but less tweeted than right-leaning videos. The attention unfolds quicker on left-leaning videos, but spans a longer period of time for right-leaning videos. Our network analysis on the early adopters and tweet cascades show that the diffusion for left-leaning videos tends to involve centralized actors, while the tweet cascades for right-leaning videos start earlier in the attention lifecycle. Our findings go beyond the static picture of ideological asymmetries in digital space and provide a set of methods to quantify attention dynamics across different social platforms.

9. **Cross-partisan discussions on YouTube: Conservatives talk to liberals but liberals don't talk to conservatives**
   Wu and Resnick. *ICWSM '21*. Full paper, **best paper finalist**
   Abstract: We present the first large-scale measurement study of cross-partisan discussions between liberals and conservatives on YouTube, based on a dataset of 274,241 political videos from 973 channels of US partisan media and 134M comments from 9.3M users over eight months in 2020. Contrary to a simple narrative of echo chambers, we find a surprising amount of cross-talk: most users with at least 10 comments posted at least once on both left-leaning and right-leaning YouTube channels. Cross-talk, however, was not symmetric. Based on the user leaning predicted by a hierarchical attention model, we find that conservatives were much more likely to comment on left-leaning videos than liberals on right-leaning videos. Secondly, YouTube's comment sorting algorithm made cross-partisan comments modestly less visible; for example, comments from conservatives made up 26.3% of all comments on left-leaning videos but just over 20% of the comments were in the top 20 positions. Lastly, using Perspective API's toxicity score as a measure of quality, we find that conservatives were not significantly more toxic than liberals when users directly commented on the content of videos. However, when users replied to comments from other users, we find that cross-partisan replies were more toxic than co-partisan replies on both left-leaning and right-leaning videos, with cross-partisan replies being especially toxic on the replier's home turf.

8. **AttentionFlow: Visualising influence in networks of time series**
   Shin*, Tran*, **Wu***, Mathews, Wang, Lyall, and Xie. *WSDM '21*. Demo paper
   Abstract: The collective attention on online items such as web pages, search terms, and videos reflects trends that are of social, cultural, and economic interest. Moreover, attention trends of different items exhibit mutual influence via mechanisms such as hyperlinks or recommendations. Many visualisation tools exist for time series, network evolution, or network influence; however, few systems connect all three. In this work, we present AttentionFlow, a new system to visualise networks of time series and the dynamic influence they have on one another. Centred around an ego node, our system simultaneously presents the time series on each node using two visual encodings: a tree ring for an overview and a line chart for details. AttentionFlow supports interactions such as overlaying time series of influence, and filtering neighbours by time or flux. We demonstrate AttentionFlow using two real-world datasets, VevoMusic and WikiTraffic. We show that attention spikes in songs can be explained by external events such as major awards, or changes in the network such as the release of a new song. Separate case studies also demonstrate how an artist's influence changes over their career, and that correlated Wikipedia traffic is driven by cultural interests. More broadly,

AttentionFlow can be generalised to visualise networks of time series on physical infrastructures such as road networks, or natural phenomena such as weather and geological measurements.

7. **Unsupervised cyberbullying detection via time-informed Gaussian mixture model**
   Cheng, Shu, **Wu**, Silva, Hall, and Liu. *CIKM '20*. Full paper
   Abstract: Social media is a vital means for information-sharing due to its easy access, low cost, and fast dissemination characteristics. However, increases in social media usage have corresponded with a rise in the prevalence of cyberbullying. Most existing cyberbullying detection methods are supervised and, thus, have two key drawbacks: (1) The data labeling process is often time-consuming and labor-intensive; (2) Current labeling guidelines may not be generalized to future instances because of different language usage and evolving social networks. To address these limitations, this work introduces a principled approach for unsupervised cyberbullying detection. The proposed model consists of two main components: (1) A representation learning network that encodes the social media session by exploiting multi-modal features, e.g., text, network, and time. (2) A multi-task learning network that simultaneously fits the comment inter-arrival times and estimates the bullying likelihood based on a Gaussian Mixture Model. The proposed model jointly optimizes the parameters of both components to overcome the shortcomings of decoupled training. Our core contribution is an unsupervised cyberbullying detection model that not only experimentally outperforms the state-of-the-art unsupervised models, but also achieves competitive performance compared to supervised models.

6. **Variation across scales: Measurement fidelity under Twitter data sampling**
   Wu, Rizoiu, and Xie. *ICWSM '20*. Full paper
   Abstract: A comprehensive understanding of data quality is the cornerstone of measurement studies in social media research. This paper presents in-depth measurements on the effects of Twitter data sampling across different timescales and different subjects (entities, networks, and cascades). By constructing complete tweet streams, we show that Twitter rate limit message is an accurate indicator for the volume of missing tweets. Sampling also differs significantly across timescales. While the hourly sampling rate is influenced by the diurnal rhythm in different time zones, the millisecond level sampling is heavily affected by the implementation choices. For Twitter entities such as users, we find the Bernoulli process with a uniform rate approximates the empirical distributions well. It also allows us to estimate the true ranking with the observed sample data. For networks on Twitter, their structures are altered significantly and some components are more likely to be preserved. For retweet cascades, we observe changes in distributions of tweet inter-arrival time and user influence, which will affect models that rely on these features. This work calls attention to noises and potential biases in social data, and provides a few tools to measure Twitter sampling effects.

5. **Estimating attention flow in online video networks**
   Wu, Rizoiu, and Xie. *CSCW '19*. Full paper, **best paper honorable mention**
   Abstract: Online videos have shown tremendous increase in Internet traffic. Most video hosting sites implement recommender systems, which connect the videos into a directed network and conceptually act as a source of pathways for users to navigate. At present, little is known about how human attention is allocated over such large-scale networks, and about the impacts of the recommender systems. In this paper, we first construct the Vevo network --- a YouTube video network with 60,740 music videos interconnected by the recommendation links, and we collect their associated viewing dynamics. This results in a total of 310 million views every day over a period of 9 weeks. Next, we present large-scale measurements that connect the structure of the recommendation network and the video attention dynamics. We use the bow-tie structure to characterize the Vevo network and we find that its core component (23.1% of the videos), which occupies

most of the attention (82.6% of the views), is made out of videos that are mainly recommended among themselves. This is indicative of the links between video recommendation and the inequality of attention allocation. Finally, we address the task of estimating the attention flow in the video recommendation network. We propose a model that accounts for the network effects for predicting video popularity, and we show it consistently outperforms the baselines. This model also identifies a group of artists gaining attention because of the recommendation network. Altogether, our observations and our models provide a new set of tools to better understand the impacts of recommender systems on collective social attention.

4. **How is attention allocated? Data-driven studies of popularity and engagement in online videos**
   Wu. *WSDM '19*. Doctoral consortium
   Abstract: The share of videos on Internet traffic has been growing, e.g., people are now spending a billion hours watching YouTube videos every day. Therefore, understanding how videos capture attention on a global scale is also of growing importance for both research and practice. In online platforms, people can interact with videos in different ways -- there are behaviors of active participation (watching, commenting, and sharing) and that of passive consumption (viewing). In this paper, we take a data-driven approach to studying how human attention is allocated in online videos with respect to both active and passive behaviors. We first investigate the active interaction behaviors by proposing a novel metric to represent the aggregate user engagement on YouTube videos. We show this metric is correlated with video quality, stable over lifetime, and predictable before video's upload. Next, we extend the line of work on modelling video view counts by disentangling the effects of two dominant traffic sources -- related videos and YouTube search. Findings from this work can help content producers to create engaging videos and hosting platforms to optimize advertising strategies, recommender systems, and many more applications.

3. **Beyond views: Measuring and predicting engagement in online videos**
   Wu, Rizoiu, and Xie. *ICWSM '18*. Full paper
   Abstract: The share of videos in the internet traffic has been growing, therefore understanding how videos capture attention on a global scale is also of growing importance. Most current research focuses on modeling the number of views, but we argue that video engagement, or time spent watching is a more appropriate measure for resource allocation problems in attention, networking, and promotion activities. In this paper, we present a first large-scale measurement of video-level aggregate engagement from publicly available data streams, on a collection of 5.3 million YouTube videos published over two months in 2016. We study a set of metrics including time and the average percentage of a video watched. We define a new metric, relative engagement, that is calibrated against video properties and strongly correlate with recognized notions of quality. Moreover, we find that engagement measures of a video are stable over time, thus separating the concerns for modeling engagement and those for popularity – the latter is known to be unstable over time and driven by external promotions. We also find engagement metrics predictable from a cold-start setup, having most of its variance explained by video context, topics and channel information – R2=0.77. Our observations imply several prospective uses of engagement metrics – choosing engaging topics for video production, or promoting engaging videos in recommender systems.

2. **Will this video go viral? Explaining and predicting the popularity of YouTube videos**
   Kong, Rizoiu, **Wu**, and Xie. *WWW '18*. Demo paper
   Abstract: What makes content go viral? Which videos become popular and why others don't? Such questions have elicited significant attention from both researchers and industry, particularly in the context of online media. A range of models have been recently proposed to explain and predict popularity; however, there is a short supply of practical tools, accessible for regular users, that leverage these theoretical results. HIPie -- an

interactive visualization system -- is created to fill this gap, by enabling users to reason about the virality and the popularity of online videos. It retrieves the metadata and the past popularity series of Youtube videos, it employs the Hawkes Intensity Process, a state-of-the-art online popularity model for explaining and predicting video popularity, and it presents videos comparatively in a series of interactive plots. This system will help both content consumers and content producers in a range of data-driven inquiries, such as to comparatively analyze videos and channels, to explain and to predict future popularity, to identify viral videos, and to estimate responses to online promotion.

1. **SMASH: A cloud-based architecture for big data processing and visualization of traffic data**
   Wu, Morandini, Sinnott. *DSDIS '15*. Full paper
   Abstract: In recent times, big data has become a popular research topic and brought about a range of new challenges that must be tackled to support many commercial and research demands. The transport arena is one example that has much to benefit from big data capabilities in allowing to process voluminous amounts of data that is created in real time and in vast quantities. Tackling these big data issues requires capabilities not typically found in common Cloud platforms. This includes a distributed file system for capturing and storing data; a high performance computing engine able to process such large quantities of data; a reliable database system able to optimize the indexing and querying of the data, and geospatial capabilities to visualize the resultant analyzed data. In this paper we present SMASH, a generic and highly scalable Cloud-based architecture and its implementation that meets these many demands. We focus here specifically on the utilization of the SMASH software stack to process large scale traffic data for Adelaide and Victoria although we note that the solution can be applied to other big data processing areas. We provide performance results on SMASH and compare it with other big data solutions that have been developed.

TEACHING

- Teaching assistant in ANU graduate course COMP6490 Document Analysis (*'17*, *'18*)
- Teaching assistant in ANU undergraduate course COMP1030 Art of Computing (*'17*)

STUDENTS

- Alexander Liu (UMich PhD)
- Zain Padamsee (Bowdoin College undergrad)
- Cai Yang (ANU undergrad)

SERVICE

- ICWSM *'23* poster, demo, and dataset track co-chair
- ICWSM *'22* student volunteer and scholarship chair
- Senior program committee: ICWSM (*'22*, *'23*)
- Conference program committee/reviewer: ICWSM (*'17*, *'18*, *'21*), CSCW (*'19-'23*), WWW (*'19*, *'20*, *'23*), WebSci (*'22*, *'23*), ICIS (*'22*, *'23*), IC2S2 (*'23*), ASONAM (*'21*, *'22*), EPJ Data Science (*'21*), ACM TOIS (*'20*), AAAI (*'19*)
- Ad hoc journal reviewer: EPJ Data Science, Collective Intelligence, Social Science Computer Review, ACM TOIS

MISC.

- Organizing: co-organized the Computational Media Lab winter workshop *'19*
- Running: 19 100+km events, 60+ (ultra) marathons finisher, 2:42 marathon