Statement of Research Interests

Haewoon Kwak

“Connect” may be the best keyword of the last decades. Beyond the barriers of age, gender, race, geography, and many others, people who used to be separated are connected online. These connections are bringing about huge changes in how we create, consume, and disseminate information online. The impact of an individual can be massively and swiftly amplified by the connected others (e.g., viral retweets in Twitter) [1, 2]. This fact that an individual’s voice can reach the whole population prompts an important question: Why can’t the online space become the ideal space for democracy?

A central question of my research is rooted on the gap between the ideal and the reality. My main research theme is to develop data-driven methodologies to understand obstacles to the trusted public space online. In this context, I divide my past research into the efforts of understanding four categories of the obstacles: media bias, toxicity, polarization, and unfair representations.

Bias in media

News media, even in the era of citizen journalism, still have immense power in setting the agenda for policy-makers and the public. Unlike factuality of news, which can often be checked objectively, the bias of news can appear in various forms and is sometimes hard to detect. For instance, just the choice of what to talk about is known to have significant impact on public opinion. One way is to mitigate this issue is identifying media biases and expose them explicitly so that readers are informed when exposed to biased news.

The forms of biases can be divided into two categories: coverage bias (what to cover) and framing bias (how to cover). In journalism literature, both biases have extensively studied but with a limited set of news media, few numbers of topics, and relatively short timespan due to the scalability of coding the news items.

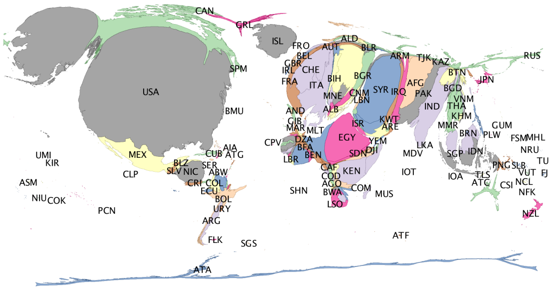

Figure 1: Biased worldview from the media in the U.S. [3]

In our first set of studies towards understanding the coverage bias of news media, which is featured in MIT Technology Review and ACM Tech News, we examined more than 195,000 events reported by media from all over the world [3]. In doing so, we not only revealed the bias of foreign news coverage but also empricially verified news value theories [4] by building a statistical model explaining the news coverage. We further revealed the difference in the attention of journalists and the public by comparing the news articles and their comments in collaboration with Al Jazeera media [5]. The stark differences between what media and public pay attention that we observed from the data naturally made us curious about its generalizability. By comparing the topics covered by news media and search terms in Google Trends, we compare media attention and public attention in large-scale across all over the world [6]. The evolution and convergence of those attentions are also studied [7, 8]. Moreover, we modeled an implicit form of coverage bias, what not to cover, as well as typical coverage bias, what to cover, through multiplex network analysis [9]. We found the hierarchy of media attention and strong regionalism in news geography, which is aligned with previous work on coverage bias, but also identified more detailed, sophisticated patterns of media and public attention, and their interactions.

The framing bias is more challenging to computationally investigate because frames are embedded in the text and thus usually implicitly conveyed. Therefore, we have been developing techniques for quantitatively characterizing frames. Using a static word embedding and a set of antonym pairs extracted from ConceptNet, we proposed a novel lightweight context-aware semantic characterization technique [10], which can measure different meanings of words according to context. It enabled us to show how some keywords (e.g., immigration or tax) are differently used between left-wing media and right-wing media. We also studied how the state-of-the-art NLP techniques can be leveraged to detect well-defined topic-agnostic news framing, such as 'Morality' or 'Public opinion.' We demonstrated that fine-tuning pretrained language models outperforms the previous SOTA for detecting news framing and implemented in our own news aggregator to help readers to understand which frame a given article uses [11].

By developing computational techniques and applying them to large-scale data, we have successfully revealed the coverage bias and framing bias of news media. Our work will provide a foundation for following studies on media bias.

Toxicity in interaction

The second obstacle I am tackling is online toxicity. While toxicity in interaction appears in offline contexts, the internet anonymity makes the problem worse, which is called online disinhibition effects [12]. Online toxicity deters healthy interactions and even brings the negative psychological consequences. We first focused on toxicity in online games. Toxicity is highly prevalent in today's online games due to its inherent competitive nature. For instance, a quarter of customer support calls to online game companies are complaints about toxic players. Victim players of toxicity are annoyed and fatigued and sometimes even leave the games. Also, the high penetration of online games to young generations makes the problem complex because toxicity experienced in youth might be connected to long term mental health issues. Thus, it is not only demanded for game companies but also, more importantly, matters for the society. The boundary of toxicity in online games is, however, unclear because the expected behavior, customs, rules, or ethics are different across games. Subjective perception of toxic playing makes toxic players themselves sometimes fail to recognize their behavior as toxic. This inherently vague nature of toxic behavior opens research challenges to define, detect, and prevent toxic behavior in a scalable manner.

From the most popular online game today, League of Legends, we collected over 10 million user reports of 1.46 million toxic players and corresponding crowdsourced decisions on whether the reported behavior is toxic or not. We built a prediction model of crowdsourced decisions based on 534 features extracted from in-game performance, user reports, and chats [14], which was widely covered by many media, including Nature News and Scientific American. Using the same dataset but with sociological theories including ingroup favoritism and outgroup hostility, we explained the toxic playing patterns across and within the team and regional differences [13]. We also studied the linguistic characteristics of toxic behavior by temporal phases of the game match [15].

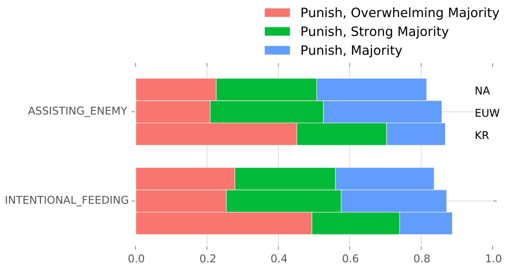

Figure 2: Regional differences in crowdsourcing decisions on toxic behavior [13]

Now we are starting to examine toxic behavior in social media. In addition to the efforts of detecting toxic comments, we aim to find toxicity triggers even before the toxic comment is written. The preliminary work demonstrates that accurate detection of toxicity triggers (and predicting upcoming toxic replies) is possible. It can bring theoretical and practical implications about understanding the origin of the toxicity and its prevention.

Polarization on social media

The primary driving forces to make “filter bubbles” online are people’s preference to connect with like-minded people and algorithms to personalize content for better engagement. Filter bubble has been identified as key culprit of social polarization that hinders the flow of ideas across communities.

We have focused on the structures and dynamics of polarized communities that emerged around Charlie Hebdo shooting in 2015 [16]. People showed an explicit endorsement of freedom of expression and freedom of the press by using the hashtag #JeSuisCharlie (“I am Charlie”), while the movement against it #JeSuisAhmed (“I am Ahmed”) and #JeNeSuisPasCharlie (“I am not Charlie”) also appear. We collected 11 million shooting-related tweets and additional 932 million tweets to construct user history and interaction networks. Using this huge collection of tweets, we then examined social factors influencing polarization based on sociological theories: clash of civilizations, density theory, and interdependence theory.

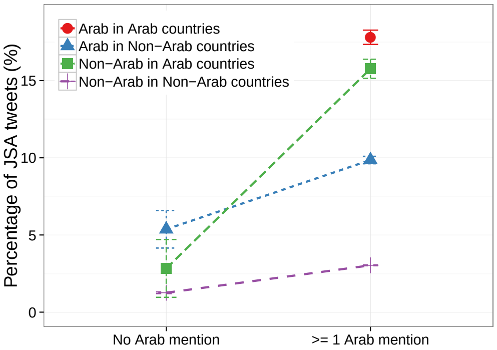

Figure 3: Impact of online and offline factors in using #JeSuisAhmed [16]

We also had an opportunity to directly compare online political discussion in homogeneous (along partisan lines) and cross-cutting (across partisan) spaces, showing the actual impact of polarization [17]. We collected 2.5M posts and 39.8M comments from four different subreddits to model homogeneous and cross-cutting spaces. Through analyses of interaction structure and linguistic patterns, we pointed to a complicated picture of online political discourse.

Unfair representations of certain demographic groups

In a multicultural society, integration and cooperation become more and more important, depending on the perception of self and other groups. One of the notable channels that affect people’s perception of the other is mass media. After being repeatedly exposed, viewers’ beliefs and attitudes are shaped by the mediated images from the media. Thus, the portrayal of different demographic groups in media, particularly advertisements that are designed for repeated exposure, can play an essential role in shaping the formation of identities of those groups.

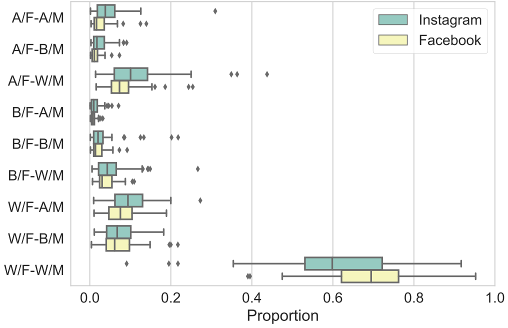

Figure 4: Demographic groups engaging in cross-sex interactions in advertising images [18]

As the first step, we conducted a large-scale analysis of the gender and racial diversity in the 85,957 advertising images of 73 international brands on Instagram and Facebook [18]. Through a comprehensive review of previous literature, we defined three metrics of the gender and racial diversity in advertisements that can be computed by automated tools: (1) how many times each gender and race appear, (2) how many times each gender and race appear in cross-sex interaction context, and (3) how many times each gender and race appear as smiling faces. Using the three metrics, our study provides a holistic view of the gender and racial diversity in today's advertising images by global brands on social media and is a great demonstration of the feasibility of our proposed metrics computed in an automated way.

Future directions

My research goal is to understand the obstacles to trusted public space online, develop methodologies to make them transparent, build frameworks to monitor them at large-scale in real-time, and make the public space online more credible. This research goal can be achieved only by the rich interdisciplinary approaches, and I believe that [XXX school] is the ideal environment for pursuing the goal of my research.

My short-term goals for future research are to go deeper into each of the aforementioned research areas. One possible way is to apply developed technologies to new media, games, networks, and issues. It quickly enables us to get insights about new environments and to shape the specific need for more appropriate techniques for them. For example, recent studies on Gab revealed its unique characteristics that attract alt-right users, conspiracy theorists, and other trolls [19]. The other way is to develop improved computational methods to handle non-textual contents. As the internet is moving to video, developing appropriate methodologies to handle them becomes more important. Below are more detailed plans for each area:

Media bias In previous work, we have analyzed textual content of news articles including headlines and body text. We plan to extend the analysis to multimedia content, such as photos and videos. As with advances of image analyses, it is starting to be possible to detect facial expressions, angles, stance, actions, or even roles of persons in the images. Using those models or fine-tuning them with domain-specific data, it is not impossible to analyze multimedia contents and examine media bias therein.

Another direction is developing methodologies to measure the impact of media bias to the public. While we showed how media attention interact with public attention, it has not been fully explored with various bias, topics, or platforms. This direction is strongly coupled with the user-level sharing on social media. I will leverage my expertise on complex networks and social media.

Toxicity in interaction The trend of penetrating mobile devices to younger generations and their active use make toxicity more important societal issues. Also, several features of social media for bigger reach and live streaming make the impact of the toxicity unexpectedly huge. Careful design of the system to protect potential victims of toxicity is more demanding than ever.

Also, longitudinal studies on toxicity in multi-community services (e.g., subreddits in Reddit) is an interesting direction because each subreddit has different norms about toxicity, and Reddit users typically participate in multiple communities. How can users change their behavior over time if their participating subreddits allow different levels of toxicity? Do they become more aggressive or less? While there are some previous studies measuring the effects of banning extreme toxic subbreddits, we focus on allowable, mild toxicity and their impacts on individual behavior in the long-term. This study will add a new dimension to effective strategies on handle toxicity in multi-community services.

Polarization on social media In previous work, while it was studied in Reddit only, we showed that cross-cutting discussion can be actively occurred in a healthy way when the appropriate environment is prepared. I plan to extend the analysis to other social media with lack of control by moderators and aim to find crucial factors for a healthy cross-cutting discussion in the wild. In parallel, I am also interested in the impact of the algorithmic bias of social media on polarization. There are a wide range of factors that social media might influence polarization, such as content recommendation, timeline optimization, or even advertisement placement.

Unfair representations As the next step of gender and racial diversity study, we plan to develop more sophisticated techniques to detect gender and race stereotyping in advertisements. For example, an advertisement that shows a female model seems to be good for gender diversity, but how about the female model cooks and a male model works at the office? Sadly, such stereotyping still appears. I plan to develop a methodology to detect more complicated forms of unfair gender and racial representations, including stereotyping and objectification.

New collaboration, new inspiration Finally, I believe that new collaborations will open a new research area that I cannot even imagine at this point. Through my past collaborations with researchers who work on other areas, such as physics, political science, social science, communication, or sports, I learned that a novel idea would come from interaction with colleagues who have a different background. [XXX school] has a huge advantage from this perspective, and I will actively seek an opportunity of collaborations.

References

[1] **H. Kwak**, C. Lee, H. Park, and S. Moon. What is Twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM, 2010.

[2] J. Reis, **H. Kwak**, J. An, J. Messias, and F. Benevenuto. Demographics of news sharing in the us twittersphere. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media*, pages 195–204. ACM, 2017.

[3] **H. Kwak** and J. An. A first look at global news coverage of disasters by using the GDELT dataset. In *International Conference on Social Informatics (SocInfo)*, pages 300–308. Springer, 2014.

[4] J. Galtung and M.H. Ruge. The structure of foreign news: The presentation of the congo, cuba and cyprus crises in four norwegian newspapers. *Journal of peace research*, 2(1):64–90, 1965.

[5] S. Abbar, A. Jisun, **H. Kwak**, M. Yacine, and B.-H. Javier. Consumers and suppliers: attention asymmetries. a case study of Al Jazeera's news coverage and comments. In *Computational Journalism Symposium*, volume 2, 2015.

[6] **H. Kwak**, J. An, J. Salminen, S.-G. Jung, and B. Jansen. What we read, what we search: Media attention and public attention among 193 countries. In *Proceedings of the 2018 World Wide Web Conference*, pages 893–902. International World Wide Web Conferences Steering Committee, 2018.

[7] J. An and **H. Kwak**. What gets media attention and how media attention evolves over time: large-scale empirical evidence from 196 countries. In *Eleventh International AAAI Conference on Web and Social Media (ICWSM)*, 2017.

[8] J. An, H. Aldarbesti, and **H. Kwak**. Convergence of media attention across 129 countries. In *International Conference on Social Informatics (SocInfo)*, pages 159–168. Springer, 2017.

[9] **H. Kwak** and J. An. Multiplex media attention and disregard network among 129 countries. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) 2017*, pages 1225–1232. ACM, 2017.

[10] J. An, **H. Kwak**, and Y.-Y. Ahn. SemAxis: A lightweight framework to characterize domain-specific word semantics beyond sentiment. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2450–2461, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[11] Y. Zhang, G.D.S. Martino, A. Barrón-Cedeño, S. Romeo, J. An, **H. Kwak**, T. Staykovski, I. Jaradat, G. Karadzhov, R. Baly, K. Darwish, J. Glass, and P. Nakov. Tanbih: Get to know what you are reading. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 223–228, Hong Kong, China, November 2019. Association for Computational Linguistics.

[12] J. Suler. The online disinhibition effect. *Cyberpsychology & behavior*, 7(3):321–326, 2004.

[13] **H. Kwak**, J. Blackburn, and S. Han. Exploring cyberbullying and other toxic behavior in team competition online games. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI)*, pages 3739–3748. ACM, 2015.

[14] J. Blackburn and **H. Kwak**. STFU NOOB!: predicting crowdsourced decisions on toxic behavior in online games. In *Proceedings of the 23rd international conference on World wide web*, pages 877–888. ACM, 2014.

[15] **H. Kwak** and J. Blackburn. Linguistic analysis of toxic behavior in an online video game. In Luca Maria Aiello and Daniel McFarland, editors, *Social Informatics*, pages 209–217, Cham, 2015. Springer International Publishing.

[16] J. An, **H. Kwak**, Y. Mejova, S.A.S. De Ogar, and B.G. Fortes. Are you Charlie or Ahmed? cultural pluralism in Charlie Hebdo response on Twitter. In *Tenth International AAAI Conference on Web and Social Media (ICWSM)*, pages 2–11, 2016.

[17] J. An, **H. Kwak**, O. Posegga, and A. Jungherr. Political discussions in homogeneous and cross-cutting communication spaces. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, volume 13, pages 68–79, 2019.

[18] J. An and **H. Kwak**. Gender and racial diversity in commercial brands' advertising images on social media. In *International Conference on Social Informatics (SocInfo)*. Springer, 2019.

[19] S. Zannettou, B. Bradlyn, E. De Cristofaro, **H. Kwak**, M. Sirivianos, G. Stringini, and J. Blackburn. What is gab: A bastion of free speech or an alt-right echo chamber. In *Companion Proceedings of the The Web Conference 2018*, pages 1007–1014. International World Wide Web Conferences Steering Committee, 2018.