Letter from the Special Issue Editors

The problem of misinformation is not new, but there is now a growing consensus that it is posing a severe threat to the well-being of societies. We do not need to look far for compelling examples—during the ongoing COVID-19 pandemic, misinformation on the coronavirus and vaccines has greatly exacerbated the public health crises across the globe. As many have observed, the harm of misinformation today has been magnified by technologies that provide easy production, access, and dissemination of information with little regard to its accuracy and credibility. The computing community must take its share of responsibilities by developing effective defenses against misinformation. This special issue is devoted to exploring how the data engineering community can contribute to this fight. While a comprehensive solution will undoubtedly require concerted effort across disciplines, we believe our community has much to offer in this fight, with our expertise on working with data at scale, with our tradition of building abstractions that enable ease of use, and with our openness to embrace collaboration and ideas from other fields.

This special issue includes six articles that provide a sample of ongoing work on combating misinformation from across the world. We start with an overview article from the University of South Florida. It examines the strengths and limitations of various existing approaches to tackling three challenges that are vital for ensuring the integrity and credibility of online information. The first challenge is about how to gauge information reliability at source level, especially at the coarse level of web domains. The second one is about fact-checking at statement level, using NLP techniques and external data such as knowledge databases. The last challenge is about other signals that can be used to estimate information credibility without examining the veracity of information content itself.

The second paper is a European collaboration on a system called *ConnectionLens* for supporting *investigative journalism*, a vital part of any modern society that is playing an increasingly important role in investigating misinformation. *ConnectionLens* helps journalists by integrating a wide range of heterogeneous, schema-less data sources into a single graph for query and exploration; it employs scalable data processing techniques to reduce the cost of constructing and searching the graph. The paper demonstrates how such a system helps with a specific use case of detecting conflict of interests—a key factor of credibility—in biomedical research; however, it benefits a broad range of tasks including fact-checking by providing the source data needed for any data-driven investigation.

Assuming we have data, the next three papers address the challenge of vetting factual claims using reference data stored in relational tables. The third paper, from Eurocom, compares four recent fact-checking systems that use tables to verify statistical claims. The comparison was carried out both analytically, focusing on how these systems differ in terms of various input and output characteristics, and empirically using several datasets of claims. These systems are based on different methodologies, ranging from natural language inference, question answering, machine learning classifiers, to text-to-SQL generation. The results can provide useful insights that help future system designers produce more advanced fact-checking systems.

The fourth paper, from Renmin University of China, has a similar goal of study but examines in more detail a narrower set of fact-checking systems—it focuses on comparing several models that are based on natural language inference. It constructs a unified framework that can be instantiated into different existing models as well as new ones. They placed a particular emphasis on allowing the framework to encode table structures in the produced language models. Their experiment results demonstrated accuracy improvement due to the inclusion of such features.

The fifth paper, a collaboration between UIC, UMich, and Google, goes beyond merely verifying the *correctness* of a claim, and detects whether it could still mislead by "cherry-picking." Specifically, the paper tackles two popular claim types of trendlines and rankings; even from the same underlying data, people can claim different trends and rankings by cherry-picking their vantage points. The paper shows how to use *perturbation* analysis to capture the robustness of claims when their vantage points are perturbed, and how to perform such analysis efficiently even over large perturbation spaces.

Finally, the sixth paper, from Cornell, outlines an end-to-end system called *WebChecker* for detecting misinformation in a very large collection of documents, e.g., from Web or social media platforms. *WebChecker* employs a wide array of fact-checking methods with different cost-accuracy trade-offs. Running an expensive deep neural net on every text snippet to detect misinformation may be more accurate, but will be prohibitively expensive at Web-scale. To look for the combination of methods that correctly detects the most amount of misinformation while staying within a cost budget, *WebChecker* adaptively switches processing methods to sample their quality and uses reinforcement learning to principally evolve its strategies.

Overall, we hope this collection of six articles together offers a sample of the ongoing work as well as open data engineering challenges in combating misinformation. Working on this special issue has been a privilege for us, as the topic has been dear to our hearts for many years. You may recall our "call to arms" to the database community in CIDR 2011—a decade later, the call is still on, and perhaps more pressing now than ever. We would like to thank the authors of this issue for their contributions, and would welcome more from the data engineering community to join this fight against misinformation.

Chengkai Li and Jun Yang
University of Texas at Arlington and Duke University, USA