



DOI:10.1145/3209581

Bias in Web data and use taints the algorithms behind Web-based applications, delivering equally biased results.

BY RICARDO BAEZA-YATES

Bias on the Web

OUR INHERENT HUMAN tendency of favoring one thing or opinion over another is reflected in every aspect of our lives, creating both latent and overt biases toward everything we see, hear, and do. Any remedy for bias must start with awareness that bias exists; for example, most mature societies raise awareness of social bias through affirmative-action programs, and, while awareness alone does not completely alleviate the problem, it helps guide us toward a solution. Bias on the Web reflects both societal and internal biases within ourselves, emerging in subtler ways. This article aims to increase awareness of the potential effects imposed on us all through bias present in Web use and content. We must thus consider and account for it in the design of Web systems that truly address people's needs.

Bias has been intrinsically embedded in culture and history since the beginning of time. However, due to

the rise of digital data, it can now spread faster than ever and reach many more people. This has caused bias in big data to become a trending and controversial topic in recent years. Minorities, especially, have felt the harmful effects of data bias when pursuing life goals, with outcomes governed primarily by algorithms, from mortgage loans to advertising personalization.²⁴ While the obstacles they face remain an important roadblock, bias affects us all, though much of the time we are unaware it exists or how it might (negatively) influence our judgment and behavior.

The Web is today's most prominent communication channel, as well as a place where our biases converge. As social media are increasingly central to daily life, they expose us to influencers we might not have encountered previously. This makes understanding and recognizing bias on the Web more essential than ever. My main goal here is thus to raise the awareness level for all Web biases. Bias awareness would help us design better Web-based systems, as well as software systems in general.

Measuring Bias

The first challenge in addressing bias is how to define and measure it. From a statistical point of view, bias is a systemic deviation caused by an inaccurate estimation or sampling process. As a result, the distribution of a variable could be biased with respect to the original, possibly unknown, distribution. In addition, cultural biases can be found in our inclinations to our shared personal beliefs, while cognitive biases affect our behavior and the ways we make decisions.

Figure 1 shows how bias influences

>> key insights

- Any remedy for bias starts with awareness of its existence.
- Bias on the Web reflects biases within ourselves, manifested in subtler ways.
- We must consider and account for bias in the design of Web-based systems that truly address the needs of users.

**YOU'RE RIGHT
AND
EVERYONE
ELSE IS
WRONG.**



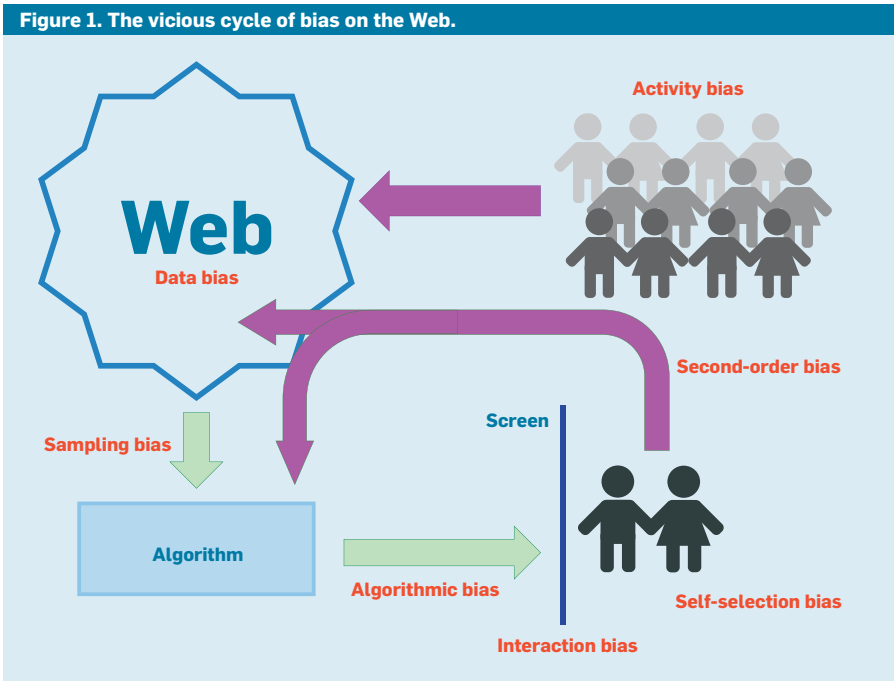
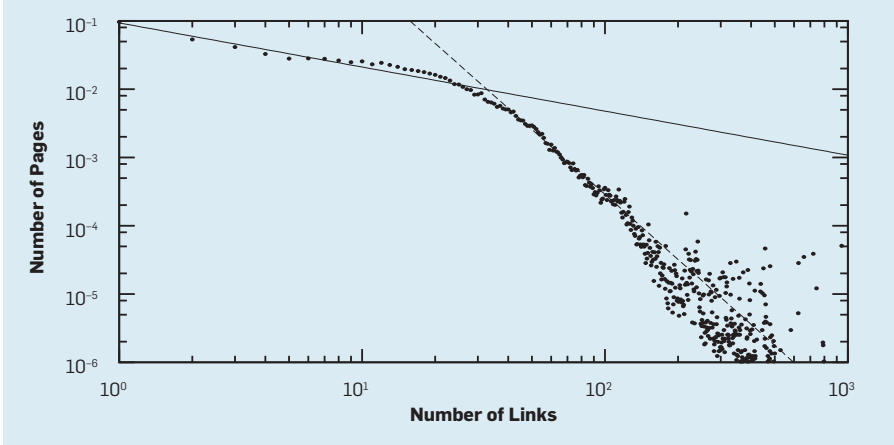


Figure 2. Shame effect (line with small trend direction) vs. minimal effort (notable trend direction) on number of links on U.K. webpages, with intersection between 12 and 13 links. Data at far right is probably due to pages having been written by software, not by Web users or developers.⁵



both the growth of the Web and its use. Here, I explain each of the biases (in red) and classify them by type, beginning with activity bias resulting from how people use the Web and the hidden bias of people without Internet access. I then address bias in Web data and how it potentially taints the algorithms that use it, followed by biases created through our interaction with websites and how content and use recycles back to the Web or to Web-based systems, creating various types of second-order bias.

Consider the following survey of research on bias on the Web, some I was involved with personally, focusing on

the significance of the categories of bias identified, not on methodological aspects of the research. For more detail, see the References and the research listed in the online appendix “Further Reading” (dl.acm.org/citation.cfm?doid=3209581&picked=formats) of this article.

Activity Bias, or Wisdom of a Few

In 2011, a study by Wu et al.²⁸ on how people followed other people on Twitter found that the 0.05% of the most popular people attracted almost 50% of all participants;²⁸ that is, half of the Twitter users in the dataset were following only a few select celebrities. I

thus asked myself: What percentage of active Web users generate half the content in a social media website? I did not, however, consider the silent majority of Web users who only watch the Web without contributing to it, which in itself is a form of self-selection bias.¹⁴ Saez-Trumper and I⁸ analyzed four datasets, and as I detail, the results surprised us.

Exploring a Facebook dataset from 2009 with almost 40,000 active users, we found 7% of them produced 50% of the posts. In a larger dataset of Amazon reviews from 2013, we found just 4% of the active users. In a very large dataset from 2011 with 12 million active Twitter users, the result was only 2%. Finally, we learned that the first version of half the entries of English Wikipedia was researched and posted by 0.04% of its registered editors, or approximately 2,000 people, indicating only a small percentage of all users contribute to the Web and the notion that it represents the wisdom of the overall crowd is an illusion.

In light of such findings,⁸ it did not make sense that just 4% of the people voluntarily write half of all the reviews in the Amazon dataset. I sensed something else is at play. A month after publication of our results, my hunch was confirmed. In October 2015, Amazon began a corporate campaign against paid fake reviews that continued in 2016 by suing almost 1,000 people accused of writing them. Our analysis⁸ also found that if we consider only the reviews that some people find helpful, the percentage decreases to 2.5%, using the positive correlation between the average helpfulness of each review according to users and a proxy of text quality. Although the example of English Wikipedia is the most biased, it represents a positive bias. The 2,000 people at the start of English Wikipedia probably triggered a snowball effect that helped Wikipedia become the vast encyclopedic resource it is today.

Zipf’s least-effort principle,²⁹ also called Zipf’s law, maintains that many people do only a little while few people do a lot, possibly helping explain a big part of activity bias. However, economic and social incentives also play a role in yielding this result. For example, Zipf’s law can be seen in most Web measures

A second set of biases is due to the interaction between different types of bias. Consider Figure 4, which plots the fraction of biographies of women in Wikipedia,¹⁶ a curve that could be explained through systemic

gender bias throughout human history.²⁵ However, an underlying factor hides a deeper bias that is revealed when looking more closely at the creation process. In the category of biographies, Wikipedia statistics

show that less than 12% of Wikipedia editors are women. In other categories, gender bias is even worse, reaching 4% in geography. On the other hand, as the percentage of all publicly reported Wikipedia female editors is just 11%, biographies actually show a small positive bias. Keep in mind these values are also biased, as not all Wikipedia editors identify their gender, and females might thus be underrepresented.

Our third source of data bias is Web spam, a well-known human-generated malicious bias that is difficult to characterize. The same applies to content (near) duplication (such as mirrored websites) that, in 2003, represented approximately 20% of static Web content.¹³

Since measuring almost any bias is difficult, its effect on prediction algorithms using machine learning are likewise difficult to understand. As Web data represents a biased sample of the population to begin with, studies based on social media may have a significant amount of error we can be sure is not uniformly distributed. For the same reason, the results of such research cannot be extrapolated to the rest of the population; consider, for example, the polling errors in the 2016 U.S. presidential election,¹⁸ though online polls predicted the outcome better than live polls. Other sources of error include biased data samples (such as due to selection bias) or samples too small for the analytical technique at hand.⁷

Algorithmic Bias and Fairness

Algorithmic bias is added by the algorithm itself and not present in the input data. If the input data is indeed biased, the output of the algorithm might also reflect the same bias. However, even if all possible biases are detected, defining how an algorithm should proceed is generally difficult, in the same way people disagree over what is a fair solution to any controversial issue. It may even require calling on a human expert to help detect if an output indeed includes any bias at all. In a 2016 research effort that used a corpus of U.S. news to learn she-he analogies through word embeddings, most of the results was reported as biased, as in nurse-surgeon and diva-superstar instead of queen-king.⁹ A quick Web search showed that approxi-

Figure 5. Heat maps of eye-tracking analysis on web-search results pages, from 2005 (left) to 2014 (right).¹⁸

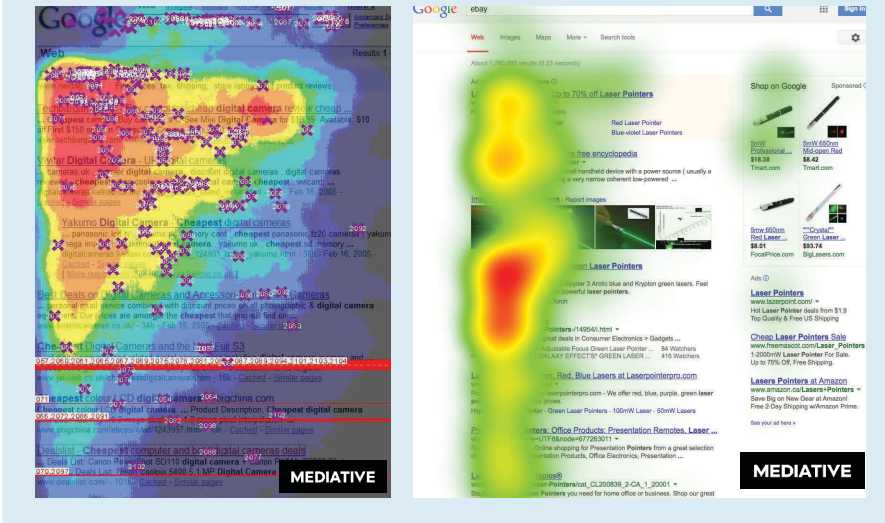
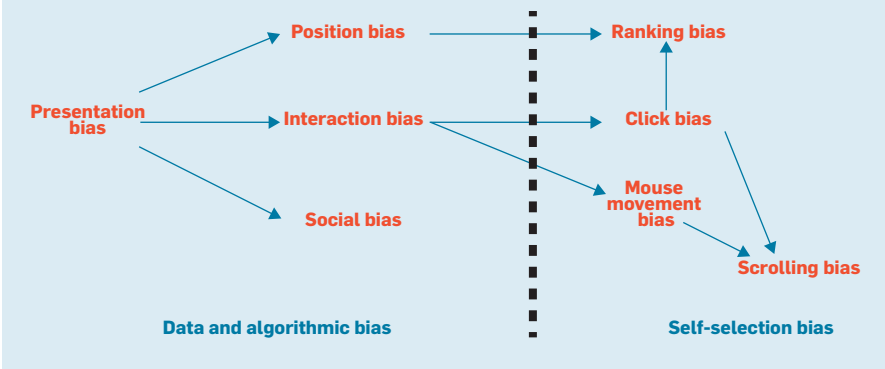


Figure 6. Dependency graph of biases affecting user interaction.



Possible classification of biases whereby the cultural and cognitive columns are user-dependent.

Bias Type	Statistical	Cultural	Cognitive
Algorithmic	●	?	?
Presentation	●		
Position	●		
Sampling	●		
Data	●	●	●
Second-order	●	●	●
Activity		●	●
User Interaction		●	●
Ranking		●	●
Social		●	●
Self-selection			●

mately 70% of influential journalists in the U.S. were men, even though at U.S. journalism schools, the gender proportions are reversed. Algorithms learning from news articles are thus learning from texts with demonstrable and systemic gender bias. Yet other research has identified the presence of other cultural and cognitive biases.^{10,22}

On the other hand, some Web developers have been able to limit bias. “De-biasing” the gender-bias issue can be addressed by factoring in the gender subspace automatically.⁹ Regarding geographical bias in news recommendations, large cities and centers of political power surely generate more news. If standard recommendation algorithms are used, the general public likely reads news from a capital city, not from the place where they live. Considering diversity and user location, Web designers can create websites that give a less centralized view that also shows local news.¹⁵

“Tag recommendations,” or recommending labels or tags for items, is an extreme example of algorithmic bias. Imagine a user interface where a user uploads a photo and adds various tags, and a tag recommendation algorithm then suggests tags that people have used in other photos based on collaborative filtering. The user chooses the ones that seem correct, enlarging the set of tags. This sounds simple, but a photo-hosting website should not include such functionality. The reason is that the algorithm needs data from people to improve, but as people use recommended tags, they add fewer tags of their own, picking from among known tags while not adding new ones. In essence, the algorithm is doing prolonged hara-kiri on itself. If we have a “folksonomy,” or tags that come only from people, websites should not themselves recommend tags. On the other hand, many websites use this idea to provide the ability to search similar images through related tags.

Another critical class of algorithmic bias in recommender systems is related to what items the system chooses to show or not show on a particular webpage. Such bias affects user interaction, as explored next. There is ample research literature on all sorts of algorithmic bias; see the online appendix for more.



In addition to the bias introduced by interaction designers, users have their own self-selection bias.



Bias on User Interaction

One significant source of bias is user interaction, not only on the Web, but from two notable sources: the user interface and the user’s own self-selected, biased interaction. The first is “presentation bias,” whereby everything seen by the user can get clicks while everything else gets no clicks. This is particularly relevant in recommendation systems. Consider a video-streaming service in which users have hundreds of recommendations they can browse, though the number is abysmally small compared to the millions that could potentially be offered. This bias directly affects new items or items that have never been seen by users, as there is no usage data for them. The most common solution is called “explore and exploit,” as in Agarwal et al.,² who studied a classical example applied to the Web. It exposes part of user traffic to new items randomly intermingled with top recommendations to explore and, if chosen, exploit usage data to reveal their true relative value. The paradox of such a solution is that exploration could imply a loss or an opportunity cost for exploiting information already known. In some cases, there is even a loss of revenue (such as from digital ads). However, the only way to learn and discover (new) good items is exploration.

“Position bias” is the second bias. Consider that in western cultures we read from top to bottom and left to right. The bias is thus to look first toward the top left corner of the screen, prompting that region to attract more eyes and clicks. “Ranking bias” is an important instance of such bias. Consider a Web search engine where results pages are listed in relevant order from top to bottom. The top-ranked result will thus attract more clicks than the others because it is both the most relevant and also ranked in the first position. To avoid ranking bias, Web developers need to de-bias click distribution so they can use click data to improve and evaluate ranking algorithms.^{11,12} Otherwise, the popular pages become even more popular.

Other biases in user interaction include those related to user-interaction design; for example, any webpage where a user needs to scroll to see additional content will reflect bias like


presentation bias. Moreover, content near images has a greater probability of being clicked, because images attract user attention. Figure 5 shows examples from eye-tracking studies whereby, after universal search (multiple types of answers) is introduced, the non-text content counteracts position bias in the results page;¹⁸ it also shows the advertising column on the right would attract additional attention.

Social bias defines how content coming from other people affects our judgment. Consider an example involving collaborative ratings: Assume we want to rate an item with a low score and see that most people have already given it a high score. We may increase our score just thinking that perhaps we are being too harsh. Such bias has been explored in the context of Amazon reviews data²⁶ and is often referred to as “social conformity,” or “the herding effect.”²⁰


Finally, the way a user interacts with any type of device is idiosyncratic. Some users are eager to click, while others move the mouse to where they look. Mouse movement is a partial proxy for gaze attention and thus a computationally inexpensive replacement for eye tracking. Some of us may not notice the scrolling bar, others prefer to read in detail, and yet others prefer just skim. In addition to the bias introduced by interaction designers, users have their own self-selection bias. White²⁷ explored a good example of how cultural and cognitive biases affect Web search engines, showing that users tend to choose answers aligned with their existing beliefs.

To make bias even more complex, interaction biases cascade through the system, and Web developers have great difficulty trying to isolate them. Figure 6 outlines an example of how such biases cascade and depend on one another, implying that Web developers are always seeing their combined effects. Likewise, users who prefer to scroll affect how they move the mouse, as well as which elements of the screen they are able to click.

Interaction biases are crucial to analyzing the user experience, as well as to a website’s overall performance, as many Web systems are optimized through implicit user feedback. As such optimized systems are increasingly based in machine learning, they learn to reinforce their own biases or the biases of other linked systems,



As any attempt to be unbiased might already be biased through our own cultural and cognitive biases, the first step is thus to be aware of bias.



yielding sub-optimal solutions and/or self-fulfilling prophecies. These systems sometimes even compete among themselves, such that an improvement in one results from degradation of another that uses a different (inversely correlated) optimization function. A classic example is the tension between improving the user experience and increasing monetization (such as the way increasing numbers of ads generally diminishes the user experience).

Vicious Cycle of Bias

Bias begets bias. Imagine we are a blogger planning our next blog post. We first search for pages about the topic we wish to cover. We then select a few sources that seem relevant to us. We select several quotes from these sources. We write new content, putting the quotes in the right places, citing the sources. And, finally, we publish the new entry on the Web.

This content-creation process does not apply solely to bloggers but also to content used in reviews, comments, social network posts, and more. The problem of drifting off message occurs when a subset of content is selected based on what the search engine being used believes is relevant. The ranking algorithm of the search engine thus biases a portion of a given topic’s organic growth on the Web. A study my colleagues and I conducted in 2008⁶ found that approximately 35% of the content on the Web in Chile was duplicated, and we could trace the genealogy of the partial (semantic) duplication of those pages. Today, the semantic-duplication effect might be even more widespread and misleading.

The process creates a vicious cycle of second-order bias, as some content providers get better rankings, leading to more clicks; that is, the rich get richer. Moreover, the duplication of content only compounds the problem of distinguishing good pages from bad pages. In turn, Web spammers make use of content from good pages to appear themselves to be quality content, only adding to the problem. So, paradoxically, search engines harm themselves unless they *do not* account for all biases.

Another example of second-order bias comes from personalization algorithms (such as the filter-bubble effect),²¹ which do not affect Web content but rather the content exposed to the

user. If a personalization algorithm uses only our interaction data, we see only what we want to see, thus biasing the content to our own selection biases, keeping us in a closed world, closed off to new items we might actually like. This issue must be counteracted through collaborative filtering or task contextualization, as well as through diversity, novelty, serendipity, and even, if requested, giving us the other side. This has a positive effect on online privacy because, by incorporating such techniques, less personal information is required.

Conclusion


The problem of bias is much more complex than I have outlined here, where I have covered only part of the problem. Indeed, the foundation involves all of our personal biases. On the contrary, many of the biases described here manifest beyond the Web ecosystem (such as in mobile devices and the Internet of Things). The table here aims to classify all the main biases against the three types of bias I mentioned earlier. We can group them in three clusters: The top one involves just algorithms; the bottom one—activity, user interaction, and self-selection—involves those that come just from people; and the middle one—data and second-order—includes those involving both. The question marks in the first line indicate that each program probably encodes the cultural and cognitive biases of their creators. One antecedent to support this claim is an interesting data-analysis experiment where 29 teams in a worldwide crowdsourcing challenge performed a statistical analysis for a problem involving racial discrimination.³

In early 2017, US-ACM published the seven properties algorithms must fulfill to achieve transparency and accountability:¹ awareness, access and redress, accountability, explanation, data provenance, auditability, and validation and testing. This article is most closely aligned with awareness. In addition, the IEEE Computer Society also in 2017 began a project to define standards in this area, and at least two new conferences on the topic were held in February 2018. My colleagues and I are also working on a website with resources on “fairness measures” related to algorithms (<http://fairness-measures.org/>), and there are surely

other such initiatives. All of them should help us define the ethics of algorithms, particularly with respect to machine learning.

As any attempt to be unbiased might already be biased through our own cultural and cognitive biases, the first step is thus to be aware of bias. Only if Web designers and developers know its existence can they address, and if possible, correct them. Otherwise, our future could be a fictitious world based on biased perceptions from which not even diversity, novelty, or serendipity would be able to rescue us.

Acknowledgments

I thank Jeanna Matthews, Leila Zia, and the anonymous reviewers for their helpful comments, as well as for Amanda Hirsch for her earlier English revision. 

References

1. ACM U.S. Public Policy Council. *Statement on Algorithmic Transparency and Accountability*. ACM, Washington, D.C., Jan. 2017; https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf
2. Agarwal, D., Chen, B.-C., and Elango, P. Explore/exploit schemes for Web content optimization. In *Proceedings of the Ninth IEEE International Conference on Data Mining* (Miami, FL, Dec. 6–9). IEEE Computer Society Press, 2009.
3. Baeza-Yates, R., Castillo, C., and López, V. Characteristics of the Web of Spain. *Cybermetrics* 9, 1 (2005), 1–41.
4. Baeza-Yates, R. and Castillo, C. Relationship between Web links and trade (poster). In *Proceedings of the 15th International Conference on the World Wide Web* (Edinburgh, U.K., May 23–26). ACM Press, New York, 2006, 927–928.
5. Baeza-Yates, R., Castillo, C., and Efthimiadis, E.N. Characterization of national Web domains. *ACM Transactions on Internet Technology* 7, 2 (May 2007), article 9.
6. Baeza-Yates, R., Pereira, Á., and Ziviani, N. Genealogical trees on the Web: A search engine user perspective. In *Proceedings of the 17th International Conference on the World Wide Web* (Beijing, China, Apr. 21–25). ACM Press, New York, 2008, 367–376.
7. Baeza-Yates, R. Incremental sampling of query logs. In *Proceedings of the 38th ACM SIGIR Conference* (Santiago, Chile, Aug. 9–13). ACM Press, New York, 2015, 1093–1096.
8. Baeza-Yates, R. and Saez-Trumper, D. Wisdom of the crowd or wisdom of a few? An analysis of users' content generation. In *Proceedings of the 26th ACM Conference on Hypertext and Social Media* (Guzelyurt, TRNC, Cyprus, Sept. 1–4). ACM Press, New York, 2015, 69–74.
9. Bolukbasi, R., Chang, K.W., Zou, J., Saligrama, V., and Kalai, A. Man is to computer programmer as woman is to homemaker? De-biasing word embeddings. In *Proceedings of the 30th Conference on Neural Information Processing Systems* (Barcelona, Spain, Dec. 5–10). Curran Associates, Inc., Red Hook, NY, 2016, 4349–4357.
10. Caliskan, A., Bryson, J.J., and Narayanan, A. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (Apr. 2017), 183–186.
11. Chapelle, O. and Zhang, Y. A dynamic Bayesian network click model for Web search ranking. In *Proceedings of the 18th International Conference on the World Wide Web* (Madrid, Spain, Apr. 20–24). ACM Press, New York, 2009, 1–10.
12. Dupret, G.E. and Piwowarski, B. A user-browsing model to predict search engine click data from past observations. In *Proceedings of the 31st ACM SIGIR Conference* (Singapore, July 20–24). ACM Press, New York, 2008, 331–338.

13. Fetterly, D., Manasse, M., and Najork, M. On the evolution of clusters of near-duplicate webpages. *Journal of Web Engineering* 2, 4 (Oct. 2003), 228–246.
14. Gong, W., Lim, E.-P., and Zhu, F. Characterizing silent users in social media communities. In *Proceedings of the Ninth International AAAI Conference on Web and Social Media* (Oxford, U.K., May 26–29). AAAI, Fremont, CA, 2015, 140–149.
15. Graells-Garrido, E. and Lalmas, M. Balancing diversity to countermeasure geographical centralization in microblogging platforms. In *Proceedings of the 25th ACM Conference on Hypertext and Social Media* (Santiago, Chile, Sept. 1–4). ACM Press, New York, 2014, 231–236.
16. Graells-Garrido, E., Lalmas, M., and Menczer, F. First women, second sex: Gender bias in Wikipedia. In *Proceedings of the 26th ACM Conference on Hypertext and Social Media* (Guzelyurt, TRNC, Cyprus, Sept. 1–4). ACM Press, New York, 2015, 165–174.
17. Lazer, D.M.J. et al. The science of fake news. *Science* 359, 6380 (Mar. 2018), 1094–1096.
18. Mediative. *The Evolution of Google's Search Results Pages & Effects on User Behaviour*. White paper, 2014; <http://www.mediative.com/SERP>
19. Mercer, A., Deane, C., and McGeeney, K. *Why 2016 Election Polls Missed Their Mark*. Pew Research Center, Washington, D.C., Nov. 2016; <http://www.pewresearch.org/fact-tank/2016/11/09/why-2016-election-polls-missed-their-mark/>
20. Olteanu, A., Castillo, C., Diaz, F., and Kiciman, E. *Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries*. SSRN, Rochester, NY, Dec. 20, 2016; <https://ssrn.com/abstract=2886526>
21. Pariser, E. *The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think*. Penguin, London, U.K., 2011.
22. Saez-Trumper, D., Castillo, C., and Lalmas, M. Social media news communities: Gatekeeping, coverage, and statement bias. In *Proceedings of the ACM International Conference on Information and Knowledge Management* (San Francisco, CA, Oct. 27–Nov. 1). ACM Press, New York, 2013, 1679–1684.
23. Silberzahn, R. and Uhlmann, E.L. Crowdsourced research: Many hands make tight work. *Nature* 526, 7572 (Oct. 2015), 189–191; <https://psyarxiv.com/qkwst/>
24. Smith, M., Patil, D.J., and Muñoz, C. *Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights*. Executive Office of the President, Washington, D.C., 2016; https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf
25. Wagner, C., Garcia, D., Jadidi, M., and Strohmaier, M. It's a man's Wikipedia? Assessing gender inequality in an online encyclopedia. In *Proceedings of the Ninth International AAAI Conference on Web and Social Media* (Oxford, U.K., May 26–29). AAAI, Fremont, CA, 2015, 454–463.
26. Wang, T. and Wang, D. Why Amazon's ratings might mislead you: The story of herding effects. *Big Data* 2, 4 (Dec. 2014), 196–204.
27. White, R. Beliefs and biases in Web search. In *Proceedings of the 36th ACM SIGIR Conference* (Dublin, Ireland, July 28–Aug. 1). ACM Press, New York, 2013, 3–12.
28. Wu, S., Hofman, J.M., Mason, W.A., and Watts, D.J. Who says what to whom on Twitter. In *Proceedings of the 20th International Conference on the World Wide Web* (Hyderabad, India, Mar. 28–Apr. 1). ACM Press, New York, 2011, 705–714.
29. Zipf, G.K. *Human Behavior and the Principle of Least Effort*. Addison-Wesley Press, Cambridge, MA, 1949.

Ricardo Baeza-Yates (rbaeza@acm.org) is Chief Technology Officer of NTENT, a search technology company based in Carlsbad, CA, USA, and Director of Computer Science Programs at Northeastern University, Silicon Valley campus, San Jose, CA, USA.

Copyright held by owner/author.
Publication rights licensed to ACM. \$15.00.



Watch the author discuss his work in this exclusive *Communications* video. <https://cacm.acm.org/videos/bias-and-the-web>