Working Paper: Modeling Gender Discrimination by Audiences of Online News

J. Nathan Matias
Microsoft Research
natematias@gmail.com

Hanna Wallach
Microsoft Research
hanna@dirichlet.net

ABSTRACT

The representation of women in public discourse—where they have historically been a minority—is important for fair, democratic societies. Although digital publishing has been heralded as a source of greater equality in women's representation, it also creates opportunities for new forms of discrimination, e.g., from audiences on social media. In this working paper, we evaluate the hypothesis that online news audiences on social media like, share, and reshare articles by men and women at different rates. We fit three Poisson regression models that predict social media impressions (counts of likes, shares, and reshares) using a sample of 156,523 articles published by the Daily Mail, Guardian, and Telegraph from July 1, 2011 to June 30, 2012. Our models suggest that audiences like, share, and reshare articles by men and women differently. We explore these preliminary results and highlight one newspaper section where articles by women have an incidence rate of social media impressions that is 33% of the rate for articles by men. Our preliminary findings raise questions for further research on modeling gender discrimination by online audiences.

INTRODUCTION

The representation of women in public discourse is important for equal participation within democratic societies. For example, global studies have shown that cultural attitudes toward gender equality are a central element of democratization [14]. Media coverage of women is linked with political participation; when women take visible roles in politics, more women demonstrate political knowledge and vote [9]. Female role models also influence adolescents' career decisions [32].

Although the representation of women in the news has increased over the past decade [19], gender inequality persists at the fundamental level of employment in news organizations. In the United States, for example, the journalism industry has failed to meet its own diversity hiring goals [1]. The percentage of women in US newsrooms has remained at 37% for the last 15 years [15], and the industry has maintained a trend of white male predominance that persists despite women outnumbering men in journalism schools since the 1980s [6].

Women have used the Internet to circumvent historical disparities, with parenting and feminist blogs gaining substantial visibility and power [18, 5, 24, 31]. Online publishing is inexpensive, the pool of voices is diverse, and institutional gatekeepers cannot prevent readers from accessing those voices [28]. For these reasons, proponents of online publishing have argued that by allowing citizen journalists and audiences to circumvent male-dominated institutions, online publishing broadens public conversation, making marginalized voices heard.

Despite early hopes that the Internet might foster peace [10] and global understanding [33], a growing literature has observed the reproduction and perhaps expansion of gender disparities, sexism, racism, and oligarchy among creators of online content, most notably in open source software development [27], peer production [16], news comments [26], and the videogame industry [21]. However, debates on inequities of attention and content sharing among audiences have primarily focused on concerns of political echo chambers [30] and filter bubbles [25] rather than problems of prejudice and inequality.

In this working paper,[1] we model gender discrimination by audiences of online news and provide preliminary results. Using social media impressions—i.e., counts of shares, likes, and reshares across several platforms—as our dependent variable, we test the hypothesis that online news audiences share, like, and reshare articles authored by men and women differently. We carry out this preliminary analysis using three Poisson regression models for articles published by three UK news outlets from July, 2011 through the end of June, 2012. Finally, we provide an exploratory discussion of our preliminary results.

MODELING DISCRIMINATION

Quantitative research on inequality differentiates between *discrimination* and *bias*. In economics, research on discrimination focuses on situations where "members of a minority [or other marginalized group] are treated differently (less favorably) than members of a majority group with identical productive characteristics" [3], offering no account of the beliefs or attitudes involved in discrimination [7]. Conversely, research on prejudice and bias focuses on measuring and explaining the reasons for behaviors that produce discrimination, often through social psychology and psychometrics methods [23]. Here, we explore differences between the rates that online news audiences like, share, and reshare articles by men and women. We do not discuss the reasons for these differences, focusing on discrimination rather than prejudice or bias.

DATA COLLECTION

Our data set includes 314,771 articles published online by the Guardian, Telegraph, and Daily Mail newspapers from July 1, 2011 to June 30, 2012. We obtained 143,515 Guardian articles through the Guardian OpenPlatform API. We scraped 110,029 Telegraph articles and 61,228 Daily Mail articles from their websites' daily archive pages. For the Guardian, we extracted metadata, including URLs, bylines, dates, sections, and titles

[1] Since this working paper describes work that is currently in progress, please do not cite it without first contacting the authors.

from the Guardian API. For the other two newspapers, we extracted metadata from article URLs and page contents.

We obtained the number of of likes, shares, and reshares for each article by querying Facebook, Twitter, and Google Plus in August 2012, at least one month after the publication of every article in our data set. We made these queries using the Mozilla Amo social media query system.[2] We refer to the total number of counts (i.e., likes plus shares plus reshares) for each article as that article's *social media impressions*.

We obtained byline gender for each article by extracting names from each article's byline and then coding these names for gender using automated techniques based on UK birth records [20, 11]. Our techniques are similar to those used in other quantitative studies of gender disparities online [26]. We labeled each byline as *male* if only male-identified names were present, *female* if only female-identified names were present, and *both* if men and women appeared as co-authors of the article. If a byline contained no author names (e.g., "Associated Press"), or where gender could not be identified using our automated techniques, we labeled it as *unknown*.

We obtained Guardian article sections from the Guardian API. We obtained Daily Mail and Telegraph article sections from topic designations in article URLs. For comparability across outlets, we coded sections (in consultation with multiple UK journalists) into a scheme that consists of nine categories: *arts/culture*, *entertainment*, *lifestyle*, *money/finance*, *news*, *opinion*, *science/technology*, *special audience*,[3] and *sport*.

In the rest of this working paper, we focus on a subset of 156,523 articles—26,340 Daily Mail articles, 69,597 Guardian articles, and 60,586 Telegraph articles. These articles all have bylines that were coded as either *male* or *female* and they all appeared in one the following (coded) newspaper sections: *sport*, *science/tech*, *opinion*, *news*, *money/finance*, and *lifestyle*.

MODELING SOCIAL MEDIA IMPRESSIONS

To test the hypothesis that there are byline gender differences by section in articles' social media impressions, we fit a multilevel, random-intercepts regression model for each newspaper.

Dependent Variable: Social Media Impressions

We used the articles' social media impressions as our dependent variable. Social media impressions for Daily Mail articles range from 0 to 95,350, with a mean of 129 and a median of 19. For Guardian articles, social media impressions range from 0 to 196,300, with a mean of 188 and a median of 53. Telegraph articles had social media impressions that range from 0 to 56,840, with a mean of 73 and a median of 28. Since our dependent variable is a count (i.e., positive integer), we chose to model our data using a Poisson regression framework [17].

Covariates

The covariates that we included in each of our models are listed in table 1. As well as including a byline-level binary covariate for gender and an article-level categorical covariate for newspaper section (with *news* as the reference section), we also included several other covariates, described below.

Table 1. Covariates used in all three models

Covariate	Description	Type	Covariate	Description	Type
Article-Level Covariates			Byline-Level Covariates		
X_{1ia}	log (title length)	real-valued	X_{14a}	female	binary
X_{2ia}	log (title length)^2	real-valued	X_{15a}	log (total articles)	real-valued
X_{3ia}	Tuesday	binary			
X_{4ia}	Wednesday	binary			
X_{5ia}	Thursday	binary			
X_{6ia}	Friday	binary			
X_{7ia}	Saturday	binary			
X_{8ia}	Sunday	binary	Interaction Covariates		
X_{9ia}	lifestyle	binary	X_{16ia}	female × lifestyle	binary
X_{10ia}	money/finance	binary	X_{17ia}	female × money/finance	binary
X_{11ia}	opinion	binary	X_{18ia}	female × opinion	binary
X_{12ia}	science/tech	binary	X_{19ia}	female × science/tech	binary
X_{13ia}	sport	binary	X_{20ia}	female × sport	binary

Since article titles offer key information that readers use in their decision to click on or share an article, we controlled for the order of magnitude of title length as an article-level covariate. Since we expected a nonlinear relationship, where very short and very long titles are less likely to be shared, we included this covariate in both linear and squared forms.

Journalists often report an anecdotal relationship between the day of the week on which an article is published and its popularity. We included day of the week as an article-level categorical control covariate, with Monday as the reference day.

Journalists vary in their experience and publication frequency. To control for this, we included a byline-level covariate for the total number of articles by that author in our data set.

Between-Byline Variation in Social Media Impressions

During the time period spanned by our data set, the people whose articles were published in the Guardian, Telegraph, and Daily Mail included politicians, first-time writers, television personalities, and sporting celebrities, as well as professional journalists with varying levels of experience and notability. Since some of these people are better known than others, and since our research question concerns gender differences between bylines, we fit a multilevel, random-intercepts Poisson regression model that accounts for variation between bylines.

FINDINGS

By fitting a multilevel, random-intercepts Poisson regression model for each newspaper, we found that social media impressions do differ by gender and that this difference varies with newspaper section. Our results are summarized in table 2.

In some newspaper sections, the magnitude of the difference in social media impressions by gender is very large, often favoring articles written by men. The exponential of each coefficient in a Poisson regression model is typically interpreted as an incidence rate ratio—i.e., the expected multiplicative increase in the dependent variable for a unit change in corresponding covariate, holding the other covariates constant. For the Daily Mail, an article by a woman in the sports section has an incidence rate of social media impressions that is 33% of the incidence rate for an article by a man. For the Telegraph, a news article by a woman has an incidence rate of social media impressions that is 86% of the incidence rate for an article by a man. An article in the money/finance section of the Guardian

[2] Amo was written by Cole Gillespie of Mozilla OpenNews Labs: `https://github.com/OpenNewsLabs/amo/`

[3] The *special audience* category includes commissioned articles and other content paid for by funders and corporations.

Table 2. Per-newspaper multilevel models for social media impressions.

	Dependent Variable: Social Media Impressions		
	Daily Mail	Guardian	Telegraph
Article-Level Predictors			
log (title length)	1.461***	0.615***	0.078***
	(0.024)	(0.005)	(0.007)
log (title length)^2	−0.163***	−0.161***	−0.016***
	(0.004)	(0.001)	(0.002)
Tuesday	0.149***	−0.121***	−0.056***
	(0.002)	(0.001)	(0.002)
Wednesday	0.143***	−0.068***	−0.200***
	(0.002)	(0.001)	(0.002)
Thursday	0.009***	−0.104***	−0.085***
	(0.002)	(0.001)	(0.002)
Friday	0.019***	−0.068***	0.038***
	(0.002)	(0.001)	(0.002)
Saturday	0.255***	0.170***	0.023***
	(0.003)	(0.002)	(0.002)
Sunday	0.480***	0.158***	0.162***
	(0.002)	(0.001)	(0.002)
lifestyle	0.535***	−0.043***	0.001
	(0.005)	(0.003)	(0.003)
money/finance	−2.622***	−0.186***	−0.416***
	(0.013)	(0.003)	(0.003)
opinion	−0.577***	0.261***	0.062***
	(0.008)	(0.002)	(0.003)
science & Tech	0.209***	0.541***	0.268***
	(0.003)	(0.002)	(0.003)
sport	0.836***	−0.470***	−0.181***
	(0.013)	(0.004)	(0.006)
Byline-Level Covariates			
log (total articles)	0.208***	0.100***	0.130***
	(0.029)	(0.014)	(0.016)
female	0.391***	0.089***	−0.155***
	(0.081)	(0.031)	(0.052)
Interaction Covariates			
female × lifestyle	−0.558***	0.362***	0.196***
	(0.007)	(0.005)	(0.006)
female × money/finance	−0.048**	−0.386***	−0.119***
	(0.024)	(0.005)	(0.007)
female × opinion	0.247***	−0.098***	−0.135***
	(0.024)	(0.004)	(0.007)
female × science/tech	0.130***	0.059***	−0.045***
	(0.006)	(0.003)	(0.006)
female × sport	−1.511***	−0.077***	−0.265***
	(0.035)	(0.009)	(0.012)
Random Effects			
bylines	2.992	1.76	1.5
	(1.73)	(1.327)	(1.225)
constant	0.120	3.696***	3.434***
	(0.073)	(0.024)	(0.038)
Misc.			
articles	26,340	69,597	60,586
bylines	2,092	9,403	2,933
deviance	9253390	21607226	8050010
log likelihood	−4,626,695.000	−10,803,613.000	−4,025,005.000
Akaike information criterion	9,253,434.000	21,607,271.000	8,050,054.000
Bayesian information criterion	9,253,614.000	21,607,472.000	8,050,253.000

Note: $^*p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

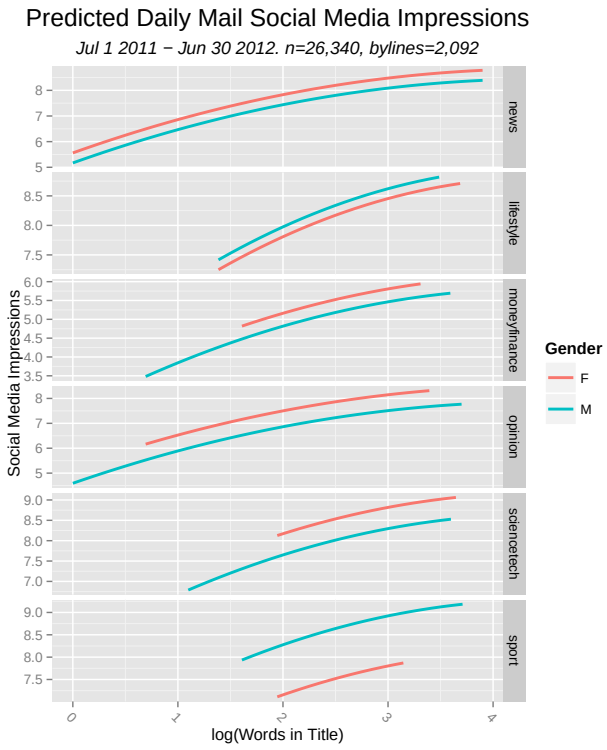

Figure 1. Multilevel, random-intercepts Poisson model predicting Daily Mail social media impressions on a Monday for a prototypical byline.

by a woman has an incidence rate of social media impressions that is 74% of the incidence rate for an article by a man. We also find the reverse relationship in some cases: lifestyle articles by women in the Telegraph and Daily Mail receive more social media impressions on average than lifestyle articles by men in the same newspaper, holding everything else constant.

To illustrate the direction of gender differences by section in each of our three models, we plotted predicted social media impressions against log-transformed title length (number of words) for articles with male and female bylines, published on Monday, with the characteristics of a prototypical byline in the corresponding newspaper (shown in figures 1, 2, and 3).

Gender Differences in Opinion Article Sharing

Opinion articles substantially influence public opinion and offer an important link in the path to funding, influence, and opportunities for elites [13]. Because opinion articles are a powerful path to opportunity, gender disparities in opinion sections contribute to wider societal disparities in opportunity for elite women. Since opinion articles are often written by people who are neither freelancers nor employees, editors have considerable flexibility to make pathways open to women.

In our data set, all three newspapers predominantly published men's opinion articles. In the Daily Mail, 21% of single-gender-identified opinion articles were by women. In the Guardian, women wrote 31% of single-gender-identified opinion articles. Finally, women's bylines accounted for just 19% of single-gender-identified opinion articles in the Telegraph.

The results in table 2 show how social media impressions differ for male and female bylines in each of the three newspapers. For the Guardian, the incidence rate of social media impressions for women's opinion articles is 99% the rate for men's articles. Meanwhile, on average, women's opinion articles in the Daily Mail have an incidence rate of social media impressions that is 190% the rate for opinion articles by men.

In the Telegraph, the small number of opinion articles by women (19%) were also shared less than men's opinion articles. On average, a woman's opinion article in the Telegraph has an incidence rate of social media impressions that is 75% of the rate of social impressions for men, holding everything else constant. These disparities in publisher and audience behavior compounded the marginalization of women already present in Telegraph opinion articles at that time. Even as the Telegraph published four times more opinion articles by men, its audiences were sharing, liking and resharing men's opinion articles much more than opinion articles by women. Perhaps knowledge of this trend motivated the Telegraph's October

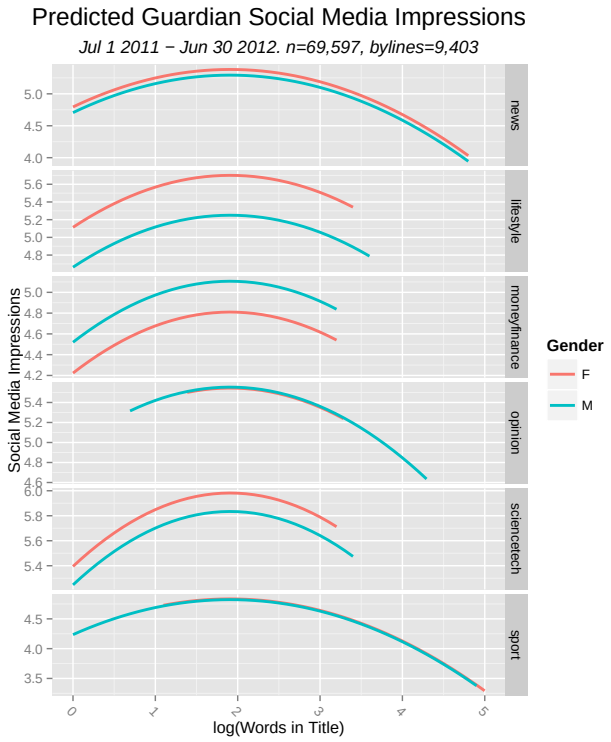

Figure 2. Multilevel, random-intercepts Poisson model predicting Guardian social media impressions on a Monday for a prototypical byline.

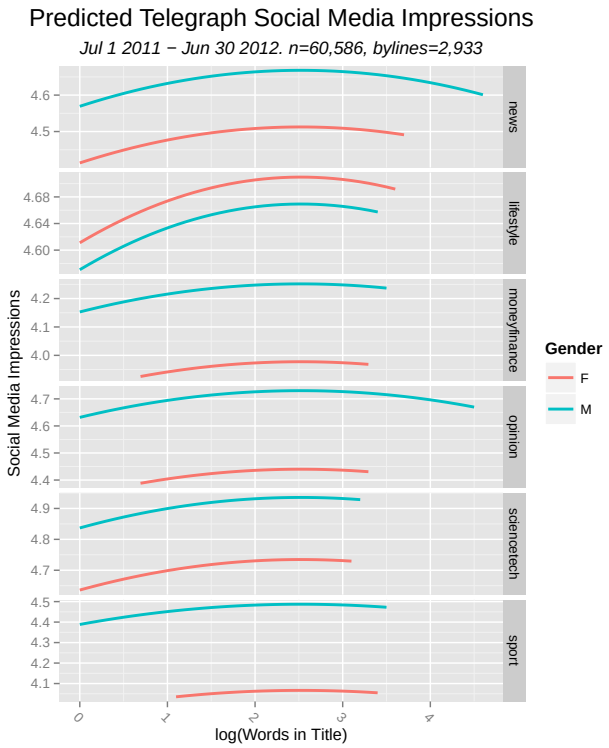

Figure 3. Multilevel, random-intercepts Poisson model predicting Telegraph social media impressions on a Monday for a prototypical byline.

2012 decision to launch "Wonder Women," an editorial unit focused on publishing and promoting women's writing [4].

Modeling Gender Discrimination by News Audiences

In this working paper, we explore an approach to modeling the contribution that audiences make to gender discrimination in who gets heard in society. Using three Poisson regression models of social media impressions, we find several examples consistent with the presence of gender discrimination by news audiences. These preliminary, work-in-progress findings identify a social factor in the problem of gender inequality that common measurements and interventions have left unaddressed: the role of online audiences in reinforcing or addressing disparities in the representation of marginalized groups, as these online audiences choose whose voices to propagate.

LIMITATIONS

Although social media platforms and online news audiences exert substantial power over whose voices receive attention [8], news organizations themselves promote content on social media. Statistical models at the New York Times have shown that promotion, including time on the homepage and sharing by official accounts, is a strong predictor of article page views [2]. Since promotion is not captured in our models, it is possible that gender differences in social media impressions are in fact related to differential promotion by news organizations. We therefore hope to account for promotion in future work.

It is possible that the content of articles produced by men and women differ in terms of the topics covered, and that this difference in fact accounts for the observed gender differences in social media impressions. For example, prior research on news articles promoted by news organizations has found that articles featuring social deviance receive more social media impressions [12]. It is also possible that authors differ in their selection of these newsworthy topics. In future work, we plan to model article content by analyzing the topic breakdown of each article (using a statistical topic model) and including the articles' topic proportions in our Poisson regression models.

Another potential limitation of our approach is that we did not account for the length of time that each article was online. In our data set, an article that was published on July 1, 2011 had 13.5 months in which to acquire social media impressions, while an article that was published on June 30, 2012 had only 1.5 months. We will address this limitation in the future.

Even if we could account for every factor within a news organization's control, we cannot attribute gender differences in social impressions to a specific issue outside of news organizations. Social media sharing is conducted in a sociotechnical context of collective action, platform affordances, and social algorithms that interact in complex ways. A substantial literature attempts to model the "interestingness" or "shareability" of specific social media posts, based on the content of those posts and the accounts that post them [22, 29]. We did not consider the content of the social media posts that mention articles or the characteristics of the accounts that share them.

FUTURE WORK

In this working paper, we presented preliminary results that identify gender discrimination by online audiences. Our findings offer a compelling argument for future work on models for studying this topic and an evaluation of alternative modeling approaches. By modeling discrimination, we hope that future work can identify the effects of online news distribution on women's representation and perhaps even establish approaches for evaluating interventions that promote equality.

ACKNOWLEDGMENTS

Lisa Evans supported data collection while at the Guardian. Lynn Cherny offered early encouragement. The MIT Center for Civic Media supported data collection and provided computational resources. This work was partially undertaken while the first author was a summer intern at Microsoft Research.

REFERENCES

1. 2015 newsroom census results. Tech. rep., American Society of Newspaper Editors, July 2015.
2. Abelson, B. How Promotion Affects Pageviews on the New York Times. *Source* (Nov. 2013).
3. Autor, D. Lecture Note: The Economics of Discrimination, Nov. 2003.
4. Barnett, E. Welcome to Wonder Women, new from The Telegraph. *Telegraph* (Oct. 2012).
5. Barnett, R. Politicians woo 'Mumsnet' generation. *BBC* (Feb. 2010).
6. Beasley, M. H., and Theus, K. T. *The new majority: a look at what the preponderance of women in journalism education means to the schools and to the professions*. University Press of America, June 1988.
7. Becker, G. S. *The economics of discrimination*. University of Chicago press, 1971.
8. Bell, E. Google and Facebook are our frenemy. Beware. *Columbia Journalism Review* (Apr. 2015).
9. Burns, N., Schlozman, K. L., and Verba, S. *The private roots of public action*. Harvard University Press, 2001.
10. Bush, V. As We May Think. *The Atlantic* (July 1945).
11. Ciot, M., Sonderegger, M., and Ruths, D. Gender Inference of Twitter Users in Non-English Contexts. In *EMNLP* (2013), 1136–1145.
12. Diakopoulos, N., and Zubiaga, A. Newsworthiness and Network Gatekeeping on Twitter: The Role of Social Deviance. In *International Conference on Weblogs and Social Media (ICWSM)* (2014).
13. Fry, E. It's 2012 already: why is opinion writing still mostly male? - Columbia Journalism Review. *Columbia Jouralism Review* (May 2012).
14. Inglehart, R., Norris, P., and Welzel, C. Gender equality and democracy. *Comparative Sociology 1*, 3 (2002), 321–345.
15. Klos, D. M. The status of women in the US media 2013. *New York: Womens Media Center* (2013).
16. Lam, S. T. K., Uduwage, A., Dong, Z., Sen, S., Musicant, D. R., Terveen, L., and Riedl, J. WP: clubhouse?: an exploration of Wikipedia's gender imbalance. In *WikiSym '11*, ACM (2011), 1–10.
17. Long, J. S. Models for count outcomes. In *Regression Models for Categorical Dependent Variables*. SAGE, 1997.
18. Lopez, L. K. The radical act of 'mommy blogging': redefining motherhood through the blogosphere. *New media & society 11*, 5 (2009), 729–747.
19. Macharia, S., O'Connor, D., and Ndangam, L. *Who Makes the News?: Global Media Monitoring Project*. World Association for Christian Communication, 2010.
20. Matias, J. N. How to Identify Gender in Datasets at Large Scales, Ethically and Responsibly | MIT Center for Civic Media, Oct. 2014.
21. Nakamura, L. Don't hate the player, hate the game: The racialization of labor in World of Warcraft. *Critical Studies in Media Communication 26*, 2 (2009), 128–144.
22. Naveed, N., Gottron, T., Kunegis, J., and Alhadi, A. C. Bad news travel fast: A content-based analysis of interestingness on twitter. In *Proceedings of the 3rd International Web Science Conference*, ACM (2011), 8.
23. Nosek, B. A., Hawkins, C. B., and Frazier, R. S. Implicit social cognition: From measures to mechanisms. *Trends in cognitive sciences 15*, 4 (2011), 152–159.
24. Nussbaum, E., and 2011. The Rebirth of the Feminist Manifesto. *NYMag.com* (Oct. 2011).
25. Pariser, E. *The filter bubble: How the new personalized web is changing what we read and how we think*. Penguin, 2011.
26. Pierson, E. Outnumbered but Well-Spoken: Female Commenters in the New York Times. In *CSCW '15*, ACM (2015), 1201–1213.
27. Reagle, J. Free as in sexist? Free culture and the gender gap. *First Monday 18*, 1 (2012).
28. Rosen, J. The people formerly known as the audience. *PressThink* (2006).
29. Suh, B., Hong, L., Pirolli, P., and Chi, E. H. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *Social computing (socialcom), 2010 ieee second international conference on*, IEEE (2010), 177–184.
30. Sunstein, C. R. *Republic. com 2.0*. Princeton University Press, 2009.
31. Valenti, V. On feminist evolutions and online revolutions. *Feministing* (2013).
32. Wolbrecht, C., and Campbell, D. E. Leading by example: Female members of parliament as political role models. *American Journal of Political Science 51*, 4 (2007), 921–939.
33. Zuckerman, E. *Rewire: Digital cosmopolitans in the age of connection*. WW Norton, Incorporated, 2013.