Who wants to read this?: A method for measuring topical representativeness in user generated content systems

Amanda Menking[1], David W. McDonald[2], Mark Zachry[2]
[1]The Information School, [2]Human Centered Design & Engineering
University of Washington
Seattle, WA, USA
{amenking, dwmc, zachry}@uw.edu

ABSTRACT
This methods paper details an approach for identifying the representativeness of content in a user generated content (UGC) system while also accounting for endogeneity bias. We leverage metadata from an independent content provider to generate sets of commercially viable terms presumed to be of interest to specific audiences linking those terms to UGC. We describe our method and heuristics at a level of detail allowing others to follow or modify it to study both content representativeness and content gaps in UGC systems. We illustrate the method by investigating how well the English language Wikipedia addresses the content interests of four sample audiences: readers of men's and women's periodicals, and readers of political periodicals geared toward either liberal or conservative ideologies. We also share preliminary findings from each case study to demonstrate our method.

Author Keywords
Methods; user generated content; representativeness; content gaps; Wikipedia

ACM Classification Keywords
H.5.3. Information interfaces and presentation (e.g., HCI): Group and Organization Interfaces -- Theory and Models. General Terms: Design, Measurement

INTRODUCTION
Assessing topical representativeness in user generated content (UGC) systems is a difficult and ambiguous task. On the one hand, content often represents the interests of contributors. Take, for example, a UGC system like Ravelry, "a knit and crochet community" dedicated to supporting users who are interested in fiber arts. While it is free to sign up for, browse, and contribute to Ravelry, one must register to access the subject-specific content the site provides. Likewise, UCG systems like Get Off My Internets (GOMI), The Potterverse, and Creepypasta have very specific content offerings—ranging from blog reviews to Harry Potter trivia and fan fiction to paranormal stories—created and consumed by users who have expressed interest in these topics and chosen to become members of these very specific UGC systems. That these UGC systems represent the interests and biases of their users would meet our expectations of those systems.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org. *CSCW '17,* February 25-March 01, 2017, Portland, OR, USA Copyright is held by the owner/author(s). Publication rights licensed to ACM. ACM 978-1-4503-4335-0/17/03…$15.00 DOI: http://dx.doi.org/10.1145/2998181.2998254

However, some UGC systems, like Wikipedia, do not require membership for users to view, edit, or contribute content. In fact, Wikipedia expresses a desire to be a general source of information *potentially* anyone and everyone finds interesting. Wikipedia also promises to be "the encyclopedia anyone can edit." Yet researchers and community members alike have recognized that many of Wikipedia's social factors affect the makeup of their contributor base [e.g., 39,42], and some of Wikipedia's policies make it more difficult for certain kinds of content to be included [51]. Moreover, studies have identified how topical skews in content increase conflict in UGC systems [29] and influence search results [28]. Our goal is to develop a systematic and reproducible method for assessing the representativeness of content in a UGC system.

BACKGROUND AND MOTIVATION
Our method is focused on understanding representativeness in proportion to specific, identifiable populations. We use *representativeness* to describe whether topical content is present in a UGC system in proportion to likely consumers of that topic. We do not use representativeness to imply aspects or the degrees to which a specific audience or population is the focus of content in a UGC system [e.g., 49]. That is, we draw a distinction between *representativeness* and *representation*, or forms of representation in a UGC system. Also, we do not use representativeness to describe how some UGC may or may not mirror societal biases and inequities. Moreover, we do not use representativeness to examine the degree to which topical content is present across a UGC system. (For example, our method does not consider the frequency of a topic and does not explicitly address the density of topical coverage, although it could be extended to address these aspects.)

Because our method considers content in relation to identifiable populations of users, our method is useful for

studying large-scale collaborative systems and not systems created or used by individuals. This approach makes the method more applicable to CSCW, social computing and other disciplines that are interested in the relationship between content, the groups who create that content, and the possible consumers of that content.

Furthermore, our method complements methods used to evaluate total coverage of topical content. *Coverage* is a measure of the degree to which some total possible content is present in the UGC system. Many studies of Wikipedia content are focused on understanding coverage or the lack of coverage (gaps in coverage). While the method we describe is intended to help understand representativeness of topical content within a given UGC system, a side effect of our method is the ability to identify both coverage and potential gaps or omissions.

Measuring Representation and Finding Gaps

Through different research methods, prior work has considered both gaps in topical content on Wikipedia—a prominent and commonly studied UGC system—and the different ways in which Wikipedia's content represents different populations. For example, Omnipedia, a system that allows users to interact with 25 different language versions of Wikipedia, visualizes content similarities and differences between language editions [4]. This approach illustrates potential gaps in content based on cultural or linguistic differences. However, mutually missing content that is a function of the participants' common biases is not detected.

Other prior work has analyzed the distribution of content into a set of content categories. Work by Kittur, Chi & Suh (2009) built upon existing work by [18,25] to identify topical distribution of articles in Wikipedia based on user generated annotations of categories. Their approach revealed the category of *popular culture* accounted for a large proportion of articles while other content categories represented very small proportions of articles. While this approach helps us understand the distribution of content, it does not tell us what content is missing, or whether the distribution of articles in the content categories effectively reveals anything about coverage of topics or the representativeness of the content. That is, this approach does not answer the question: *Who may want to read this?*

Another approach to understanding who and what is represented is to study the composition of the content. In the case of Wikipedia, researchers have studied articles to understand what they address. For example, recent studies have considered how phrases missing from Wiktionary might be identified and eventually incorporated [57], how articles about women might be expanded by pulling missing data from obituaries [36], and how German and English language Wikipedia articles about drugs might be improved by comparing them to pharmacology textbooks [31]. These approaches focus on the composition of the content and thereby inform our understanding of the representation of the object of the content.

Some recent work by Sengul-Jones identifies the challenge of focusing on identifying gaps. In her report for a Wikimedia Foundation Individual Engagement Grant entitled "Full Circle Gap Protocol," she points out that identifying the "unknown unknowns" becomes a "bestiary of gaps" [46]. She finds gaps derive from and are tied to a range of factors such as infrastructural access, skill and literacy divisions, time and interest gaps, emotion work, and knowledge-legitimacy exclusions [47]. Many of these items are specific characteristics and biases present in the range of skills and knowledge of individual contributors. In short, gaps and omissions are difficult to find, and many of the reasons why they exist may be inherent in the UGC system, a function of social and political forces, or an aspect of the users' lived experiences.

These studies help us understand coverage or gaps in Wikipedia, but many of the techniques they employ fail to provide a general method for assessing other UGC systems. Further, some of these approaches are specific to a particular topic or aspect of Wikipedia and may not apply to other UGC systems that may represent topical content differently from the way it is represented in Wikipedia. However, the challenge of assessing representativeness prompts other problems that need to be addressed first.

Identifying Topical Content: Endogenous versus Exogenous Source

Our attempt to assess content representativeness faced a number of challenges. One problem is the scope of potential topics. In some prior work, the methods to identify subsets are very narrow and do not represent the broad interests of potential content consumers. *Representativeness* would fail in that case because the population of potential consumers is so narrow as to be difficult to compare article coverage to potential readership. Studies like those of biographies [e.g., 16,52,53] or controversial issues [e.g., 26] are topically focused and provide an understanding of those topics, but cannot tell us much about *representativeness*. In other prior work, the methods generate broad subsets of content, but the potential population of content consumers is diffuse in such a way as to make judging representativeness of the content untenable. For example, very broad and diffuse readerships of content related to 'work', 'relationships' and 'leisure' [6] would make it hard to know whether these topics are reasonably represented within Wikipedia or any other UGC system.

Another underlying problem is that—within any UGC system—differential participation may yield content skews. For example, in the context of Wikipedia, a well-documented skew in participation [23] may result in a skew in topical content. Therefore, any sampling method that relies on content from Wikipedia may be unknowably biased from the beginning.

All of these concerns point to a broader challenge for techniques that purport to assay any aspect of a UGC or social media system: *the challenge of endogeneity*. Endogeneity, the state of being endogenous, or coming from within, is an underappreciated problem for many prior studies of UGC systems and social media. We will not address the problems of these studies that have clear biases based on endogenous sampling. We merely raise the issue here to acknowledge that the challenge is pervasive and that our method is one of the few methods designed to free an analysis from an endogeneity bias.

In our review of the literature, we found two notable exceptions that also made an effort to address the challenge of endogenous biases by relying on exogenous sources. Reagle & Rhue's (2011) study compares women's biographies in Wikipedia to women's biographies in the online *Encyclopædia Britannica*. This study shows the degree to which Wikipedia and *Encyclopædia Britannica* covered similar content. The other study is by Klein & Konieczny's (2015). They compare "a Wikipedia-derived gender inequality indicator (WIGI)" to four commonly used gender inequality indices. These exceptions can help expose bias in coverage that may be a function of the UGC participants' choices, but cannot tell us the degree to which the existing content is *representative* of the potential readership.

Our method attempts to address these challenges (i.e., too narrow of a scope, overly diverse potential readership, and endogeneity bias) by relying on an external content provider. In the next section, we describe how we identified potential consumer populations and how we derived content metadata that would facilitate a match to content in Wikipedia.

METHOD

The method we developed requires four steps to measure how well topical content aligns with potential readership populations (i.e., users) of a UGC system, like Wikipedia. The first step is to use an exogenous source to generate a set of unique topics of *presumed* interest to a readership population. The second step is to differentiate among identifiable groups within this population to isolate topics that are representative of the interests of these groups and not the entire population. The third step is to assess how those topics are covered in the system through the application of heuristic-guided search to resolve results that are not an exact match. Finally, the fourth step is to use an external, authoritative source to identify the size of the groups within the overall population and compare those numbers to the proportion of how their interests are represented in the UGC system.

The high-level details about how the method is executed are discussed in the following subsections. Additional details of the method and its application are then illustrated in two cases later in this paper.

Step 1: Generating Lists of Topics Associated with Specific Populations

Instead of settling for existing approaches that have some clear drawbacks (e.g., endogeneity bias, narrow topical focus), we developed a method that works independently from the unique characteristics of the UGC system under study to generate a large set of topics associated with the interests of a given population.

To generate our initial topic lists, we first identify content providers (e.g., magazines, journals, monographs, etc.) commercially targeted to a population of interest. The selection of this population must be executed with sensitivity to two of the considerations discussed earlier. First, the population must not be too narrow, identifiable only with an un-differentiatable content area. Additionally, the population must not be too broad or diffuse, making it impossible to identify meaningful groups within it that would likely have distinguishable topical interests. For the purposes of this method we call that target population a *target demographic* or, simply, *demographic*.

With the population selected, we then identify specific content sources targeted to this population, using widely available reports from the publishing industry about publications marketed to different demographics and circulation data. From this report data, we select the most widely circulated sources based on published circulation data. Additionally, we select sources based on the criteria that they were indexed uniformly. In the cases below, we used presence in the most prominent U.S.-based indexer of popular magazines, EBSCO. The selection criteria applied when using this database can be adjusted as necessary to fit different research questions (e.g., time periods, number of content sources to consider).

	jtl	year	mo	subj	ab	pu
21	Rolling Stone	2014	12	FEMINISM & music	The article presents a discussion with sing	Pe
22	Rolling Stone	2014	12	DOCUMENTARY films	The article presents a discussion with sing	Pe
23	Rolling Stone	2014	12	SINGERS	The article presents a discussion with sing	Pe
24	Rolling Stone	2014	12	HEALTH	The article presents a discussion with sing	Pe
25	Rolling Stone	2014	12	Musical Groups and Artists	The article presents a discussion with sing	Pe
26	Rolling Stone	2014	12	ARETHA Franklin Sings the Great Diva Classics	The article presents a discussion with sing	Pe
27	Rolling Stone	2014	12	RESPECT (Music)	The article presents a discussion with sing	Pe
28	Rolling Stone	2014	12	FRANKLIN, Aretha, 1942-	The article presents a discussion with sing	Pe
29	Rolling Stone	2014	12	DAVIS, Clive, 1932-	The article presents a discussion with sing	Pe
30	Rolling Stone	2014	12	MCDONALD, Audra, 1970-	The article presents a discussion with sing	Pe
31	Rolling Stone	2014	12		The article lists the top 40 music albums o	Pe
32	Rolling Stone	2014	12	POPULAR music -- 2011-2020	The article lists the top 40 music albums o	Pe
33	Rolling Stone	2014	12	SONIC Highways (Music)	The article lists the top 40 music albums o	Pe
34	Rolling Stone	2014	12	EVERYTHING Will Be Alright in the End (Music)	The article lists the top 40 music albums o	Pe
35	Rolling Stone	2014	12		The article lists the top 40 songs of 2014 a	Pe
36	Rolling Stone	2014	12	POPULAR music -- 2011-2020	The article lists the top 40 songs of 2014 a	Pe
37	Rolling Stone	2014	12	DRUNK in Love (Music)	The article lists the top 40 songs of 2014 a	Pe
38	Rolling Stone	2014	12	FRANKIE Fell in Love (Music)	The article lists the top 40 songs of 2014 a	Pe
40	Rolling Stone	2014	12	REISSUE of recorded music	The article lists the top 10 reissued music a	Pe
41	Rolling Stone	2014	12	BASEMENT Tapes Complete, The (Music)	The article lists the top 10 reissued music a	Pe
42	Rolling Stone	2014	12	BEATLES, The (Music)	The article lists the top 10 reissued music a	Pe

Figure 1. A screenshot of a database record for a content source (e.g., *Rolling Stone*) with subjects and other metadata. We derived our topics from the subject category ("subj"). Abstracting services, like EBSCO index a wide variety of content. Often these indexing records contain a wealth of potential fields that do not apply to every indexed and abstracted form of content. This screenshot shows only a few of the relevant fields for the selected sources. Similar fields (not shown) might be useful when applying this method, but when choosing a different media as the initial topical source.

Queries of the EBSCO database for the targeted publication title in a given year yielded a data file including all indexed items (articles) for the year. Each item included several fields of article-associated metadata, including, titles, subtitles, author(s), abstracts, and a list of subject terms (topical keywords). (See Figure 1.)

In our cases, we chose to use periodicals with the highest circulation numbers in the English language. For each periodical, we remove all fields except the *subject terms*. We then combine the subject terms for each periodical title for a demographic into a single list. For each of these lists, all duplicate items are removed.

These subject terms—or *keywords*—are generated by the abstracting service following traditional knowledge organization (i.e., library science) practices. The keywords are not generated by the editorial practices of the periodicals, or by us. (Detailing the methods for creating subject terms and the practices of indexing publications materials are beyond the scope of this paper and outside the bounds of the method that we define here.)

Step 2: Differentiating Groups and Associated Topics within the Overall Population

Through Step 1, the method has produced a set of items that are indicative of what topics are covered in a set of content sources for each target demographic. To facilitate an assessment of *representativeness*, though, the topic list associated with each demographic must be sorted into sub-lists. In the cases reported below, we achieved this by initially selecting periodical titles targeted to two identifiable groups maintaining separate lists of content terms for each demographic. These lists of terms could be used to assess general topical coverage for the interests of each target demographic. We can move beyond the question of coverage to assess representativeness by generating a list of topical subject terms for each demographic that is unique to that demographic. These unique lists are the disjoint sets of terms across all target demographic groups. In the demonstrations below, we compare two lists and remove all items that appear in both lists. We point out that the intersection of terms across the target demographics could also yield some interesting insights, but that is not the way we implemented our assessment of representativeness in our demonstration cases below.

Step 3: Detecting Topical Coverage in the UGC System (Wikipedia)

The next step is to apply the differentiated list of target terms to assess whether specific content exists in a given UGC system. In the demonstration cases below we applied the terms to the English language Wikipedia. The generated topic lists provide a baseline for measuring coverage of topics *presumed* to be of interest to readers from each demographic in the UGC system. The next challenge is to match topical terms to UGC to assess whether the topic from the external sources are currently present in the UGC. For our demonstrations below, we developed a set of heuristics to match topics to Wikipedia article titles.

We relied on the English language Wikipedia's native search feature as the basis for generating a possible article match. This is a likely mechanism an average user would employ. Another approach would be to use a Google search, using the features of Google to constrain the search to the domain of the specific UGC. We do note that the Google search approach might yield differential results because of the way that Google uses data (e.g., geo-location, account preferences) that it has about an individual to tailor and shape its search results. Our process involved entering each of our topics into the Wikipedia search field and evaluating the results. Based on our heuristics, we counted and tracked the search result for each topic. The topic lists include a primary topical keyword (in all caps) with some additional terms that clarify the topic. Some topics include additional parenthetical descriptors, comma series, or descriptors after dashes. When entering topics into Wikipedia's search engine, we used only top-level topic keywords and no extended descriptors.

Proper names of persons are a special case for our search and match method. The majority of abstracting services handle proper names of people in a "LASTNAME, Firstname" strategy. This is not the editorial standard for Wikipedia. The current standard in Wikipedia is "Firstname Lastname." Except, in Wikipedia, a significant number of well-known people will have a "Lastname, Firstname" redirect. In this special case, proper names of people, we modified how we entered the term metadata to align with the standards of the UGC system. That is, we swapped the "Lastname, Firstname" data from the abstracting service to align with the "Firstname Lastname" assumption of Wikipedia. The results of that search were then used.

The heuristics we used for counting matches between topics and specific Wikipedia articles are as follows:

1. If the search yielded a direct match, we recorded the name of the article page and noted it.
2. If the search resulted in a redirect, we recorded the name of the article page and noted it was the result of a redirect.
3. If the search resulted in a list of possible matches via a search results page, we carefully considered each item, with the goal of selecting the lowest ordinal result that seemed to be the closest match given additional information provided by any parenthetical descriptors from the term list, and we noted the ordinal of that result item. The research team reviewed these cases to determine how they should be counted.
4. If the search resulted in a disambiguation page, we scanned the suggestions to identify the closest match, relying on any additional information provided by any parenthetical descriptors from the term list.

5. If the search resulted in an article that was a list rather than an actual encyclopedic article, we noted it and counted it as a non-match.
6. If the search did not result in a match—either through a direct hit, a redirect, a search result selection, or a disambiguation page—we marked the term as not having a corresponding article in Wikipedia. Many searches yielded a "does not exist" message generated by Wikipedia.

Adjudicating Imprecise Matches

These heuristics were supplemented with team review discussions to adjudicate imprecise matches. Review of these imprecise matches was necessary because naming conventions in Wikipedia are not consistent, and typos occur in the databases of abstracting services. For example, if a search did not result in a match—either through a direct hit, a redirect, a search result selection, or a disambiguation page—because the metadata lacked sufficient information, such as, birth dates for individuals, we discussed the item and possible matches. In our review decisions, we erred on the side of inclusion. However, we also made exceptions if we agreed common sense challenged the results of our heuristics. For example, when a search for "Apologizing" led to a disambiguation page listing "An expression of remorse" with a hyperlink to the article "Remorse" as the first—and most applicable result—we reviewed the article, discussing it as a team, and decided "Remorse" was not a match for the keyword "Apologizing." Additionally, we did not include matches that resulted in article sub-sections rather than complete standalone articles.

Step 4: Comparing Representation of Topics in the UGC to the Population

The final step in our method is to use an external, authoritative source to identify the proportion of the target demographic within the overall population and then compare those numbers to the proportion of how their interests are covered in the UGC system. The details about how we executed this step in the two case studies are presented below.

DEMONSTRATING THE METHOD

In the following sections, we explain why we chose our two cases to test our method, frame each case with related work, present how we worked through the method in each case, and share preliminary findings as demonstrative data of our method.

Choosing the Cases

In this study, we focus on assessing the relative representativeness of content in the English language Wikipedia by applying the method described above to two cases: (1) topics perceived to be of interest to readers of "women's" and "men's" periodicals; and (2) topics perceived to be of interest to readers of "conservative" and "liberal" periodicals. We understand there is often overlap between these audiences, but we do not hypothesize any connections between the two, and—for the purpose of this study—we treat them as separate, identifiable populations.

We chose these two cases to demonstrate the method because, for both cases, (a) there are well established periodicals with targeted, commercially viable readerships; (b) there is existing work about the domain and UGC; and (c) there is some debate regarding the coverage of the domains within Wikipedia. In the case of politics, the English language Wikipedia has been portrayed as having a liberal bias [e.g., 35,48], and Conservapedia—created to counter Wikipedia—claims to be "free of liberal untruths" [14]. In the case of gender, Wikipedia has a known "gender gap" in participation [e.g., 23] and the inclusion of articles perceived to be of more interest to women is often contested [e.g., 7].

Also, for both cases, recent polls and census data for U.S. residents provide a baseline against which to measure potential representativeness as a function of a general population. The 2015 Gallup poll reports there are now more people who identify as conservatives (38% of respondents) than as moderates (34%) or liberals (24%) [45]. And the most recent U.S. Census reports there are slightly more respondents who identify as women (50.8%) than as men (49.2%) [2]. Therefore, if an "open" UGC system like Wikipedia is representative of the interests of these populations, we would expect to see distributions of topical content in relation to the possible interests of each audience.

Some may argue periodicals such as popular magazines and mass media in general are not good sources to use to elicit what is and is not a topic of interest to "men" or "women," or "conservatives" or "liberals." The common argument against these sources is that the commercial needs of popular content providers serve only to reify binaries and stereotypes. That a binary may or may not be reified in the content of the individual sources is not the focus of this analysis. Nor was our goal to examine how gender and political parties are socially constructed. Rather, our goal was to find sources external to Wikipedia that exhibit both established readership and commercial viability. Further, the established commercial viability of these sources makes claims to the interests and desirability of the topics, if not the specific presentation of the content. This means that the potential readership of a UGC might want to learn more about these topics, even if it were from a different perspective than provided in the given external source. Furthermore, because of Wikipedia's policies regarding notability [55,56], searching for keywords associated with content published in popular periodicals with high distribution is likely to reflect the potential inclusion of the associated content in Wikipedia simply because an external, verifiable source is available.

Demonstration Case 1: Political Ideologies

Recent events have drawn public attention to the ways in which UGC systems—and social media in particular—

reflect and drive political discourse, sometimes using unexpected and difficult to detect mechanisms. For example, Facebook has been accused of manipulating its "trending topics" to reflect anti-conservative bias [e.g., 34], activists claim Twitter has censored #WhichHillary, an anti-Hillary Clinton hashtag [36], and TwitterAudit suggests that 75% of Donald Trump's followers are bots rather than people [58].

In the case of Wikipedia, five years after it was established in 2001, Andrew Schlafly created Conservapedia in response to what he perceived to be the liberal biases perpetuated by the growing UGC system [1]. RationalWiki, a wiki that does not claim to be an encyclopedia or value neutral point of view, was created in 2007 as a counter to Conservapedia after some editors were banned from the latter. Though Conservapedia and RationalWiki have failed to grow at the same rate and in the same ways as Wikipedia, the contributors to these UGC systems remain active and continue to maintain that Wikipedia does not represent their interests.

Scholars have studied political biases and content coverage in traditional media for decades. More recently researchers have begun to think about how UGC reflects political ideologies and how both citizens and politicians leverage UGC systems [e.g., 21,22]. Greenstein & Zhu (2012) sought to assess the political bias of Wikipedia content relative to the ideology of the two principal U.S. political parties. They found articles are created with "Democrat leanings" but tend to become more neutral with time and additional revisions, and that the total bias in Wikipedia changes as content is added—though the slant of individual articles may not change significantly. Similarly, Kalla & Aronow (2015) looked at articles about U.S. senators to determine whether negative or positive sentiments persist, to find a bias toward positivity over time. Brown (2011) also considered existing coverage of articles related to politics, looking at both U.S. politicians' biographies and articles about U.S. election results. He found coverage was good for recent and popular topics, but that there were omissions for older and more obscure topics.

These studies about Wikipedia and politics have paid careful attention to and used various techniques to analyze the content that exists within Wikipedia, what we defined as the representation, but they have not addressed the issue of *representativeness*, especially as it relates to broad areas of interest. Our goal in this case was to use the method to test whether the English language Wikipedia reflects general representativeness of topics that are *presumed* to be of interest to an established readership of "conservative" and "liberal" periodicals.

Applying the Method

As noted above, we chose sets of periodicals targeted to specific audiences based on the highest circulation in the U.S. and accounting for whether they had been indexed uniformly by EBSCO. In our first case, we chose periodicals targeted to "conservative" and "liberal" readers. (See Table 1.) There is one notable aspect of our method that arises here. The publication *Cosmopolitan* is a very high circulation magazine that identifies its target readership as both "women" and "liberal." In building our example cases, we decided not to include it here and instead used *Cosmopolitan* for our second demonstration below. We made this choice because the primary audience is women. We recognize political ideology and gender are not separable, but as a demonstration this separation was a clearer way to illustrate the method. We replaced *Cosmopolitan* with the periodical that had the next highest circulation in the U.S.

Periodicals Marketed to Liberals	**Periodicals Marketed to Conservatives**
Mother Jones	*American Spectator*
The Nation	*National Review*
The New Republic	*Reason*
Rolling Stone	*Saturday Evening Post*

Table 1. Periodicals with the highest subscribed circulation in the U.S. by targeted readership (i.e., category).

Following the method described above, we removed all fields except the subject terms for one year's worth (2014) of metadata for each periodical. We then combined the subject terms for each of the four periodical titles in a category (e.g., "conservative") into a single list. This resulted in 4,576 terms from "conservative" periodicals and 7,403 terms from "liberal" periodicals. For each of these lists, all duplicate items were removed, leaving 3,313 unique terms from "conservative" periodicals and 6,140 unique terms from "liberal" periodicals. This gave us one list of subject terms (or keywords) for 2014 from the four "liberal" periodicals listed above and a separate list for subject terms for 2014 from the four "conservative" periodicals listed above.

We generated two sets of disjoint topic terms by removing all terms that were common to the set of terms (keywords) between the "liberal" list and the "conservative" list.

We then used a random number generator to select 400 terms from each category list and searched for the resulting 800 terms using Wikipedia's native search feature. We recorded what we found using the heuristics described above for each search term and met to adjudicate imprecise matches. At times, there were interesting instances that generated no matches until we slightly changed the entry of the metadata. For example, "DUOLINGO Inc." did not result in any matches, but "Duolingo" resulted in a direct hit. These results were put into a category of "Direct hit edited term" to differentiate them from a clear direct match.

Demonstrative Data

We found that 73.8% of the randomly selected 400 keywords from conservative-oriented periodicals were covered, and 81.5% of the randomly selected 400 keywords from liberal-oriented periodicals were covered. This represents a 7.7% difference in topical coverage. Overall, 18.5% of the randomly selected topics from periodicals targeted to "liberal" readers and 26.3% of the randomly selected topics from periodicals targeted to "conservative" readers did not have corresponding articles. (See Figure 2.)

	Topics from Sources Marketed to Liberal Perspective		Topics from Sources Marketed to Conservative Perspective	
	count	%	count	%
Does not exist	67	18.51%	100	26.25%
Direct match	179	49.45%	144	37.80%
Direct hit edited term	6	1.66%	1	0.26%
Redirect	73	20.17%	80	21.00%
1st search result	7	1.93%	18	4.72%
2nd search result	0	0.00%	3	0.79%
Nth search result	8	2.21%	17	4.46%
Applied multiple rules	12	3.31%	7	1.84%
Match disambiguation	10	2.76%	11	2.89%
Compound keyword	8		2	
List	2		3	
Vague/Generic Term	24		11	
Duplicate/Variant	0		0	
Other disqualification	4		3	
Total terms	400		400	
Terms for analysis	362		381	
Term coverage		81.49%		73.75%

Figure 2. Summary of results for topics from sources marketed to "liberal" and "conservative" readers. The last 5 gray shaded rows were omitted from the analysis.

The "direct match" heuristic was the most activated heuristic in our study with a total of 144 terms from the list taken from "conservative" periodicals and 179 terms from the list taken from "liberal" periodicals having direct hits—or corresponding articles—when entered into Wikipedia's search feature. Alternately, another way to consider representativeness is to note that only 67 keywords from the list from "liberal" periodicals did not have corresponding articles, whereas 100 keywords from the list from "conservative" periodicals did not have corresponding articles.

In Figure 2, the bottom five shaded rows represent the number of terms excluded from the analysis. The number of topics excluded from analysis (e.g., because they matched something that was not a legitimate article, because the topic could not be effectively disambiguated, or the topic was conceptually vague without access to the original material, etc.) favored the inclusion of content represented by keywords taken from the "conservative" list. We excluded twice as many keywords from the list taken from "liberal" magazines (38 were excluded) than from "conservative" magazines (19 were excluded). The majority of excluded terms were either too vague (e.g., "PROGRESS"), or compound keywords (e.g., "DRUGS & crime").

Summary

Our preliminary findings suggest the English language Wikipedia is more likely to include topics presumed to be of interest to readers of "liberal" periodicals. In the case of political ideology, then, there is a clear need to investigate biases and gaps, and our method provides a systematic approach of identifying articles that might be of interest to "liberal" and "conservative" readers.

Demonstration Case 2: The "Gender Gap"

Recent UGC studies have begun to unpack the complex implications of participation and content differences among the genders. For example, women who contribute reviews to the Internet Movie Database (IMDb) enjoy less prestige and smaller audiences, even when they adjust their communication styles to mimic those of men who contribute to the same site [20]. Similarly, Facebook content perceived as "male" receives more feedback, even when posted by users who identify as women [34], and Google image searches for many occupations result in both gender stereotypes and underrepresentation of women [28].

To investigate the potential gender valence of content, prior studies have analyzed conversation topics of passing strangers [6,40], online comments on *New York Times* articles [41], tweets about climate change [26], and images curated via Pinterest [10,38] to determine the kinds of topics that may be of interest to men or women. These studies generally rely on an inductive approach: first identifying the gender (or, in some cases, biological sex) of the contributor and then classifying the kinds of content he or she contributed. Although this inductive approach indicates the types of content contributed, it cannot effectively speak to the *representativeness* of the total amount of content in the system.

Another approach [11] relies upon topics selected by the researchers to test how participants respond to gender stereotypes. While this method represents one approach, it also presents methodological challenges as it generates results that can be difficult to reproduce, and—as a method—it is likely to introduce biases from the idiosyncrasies of the researchers.

In the case of Wikipedia, research [3,12,15,23,32,33] has identified differential participation in contributors who identify as men or as women, and the phenomenon has come to be known as the "gender gap." In 2015 the Wikimedia Foundation launched the Inspire Campaign with

the goal of funding ideas that address the "gender gap." One idea creator wrote, "I see a lot of speculation here about whether the average potential female editor of Wikipedia is interested in fashion [...] or in female scientists [...]. It is hard to tailor our recruitment without having more data" [50]. A different but similar proposal noted, "There is a lack of clarity over what subjects women are interested in, what articles they edit, whether more women on Wikipedia would mean more coverage of certain areas, etc. We should research this rather than guessing" [44]. Although neither idea generated a fundable proposal, both touched on a challenging question. They also troubled the implicit assumption that a diversity of contributors in an "open" UGC system will ensure *representative* content.

Direct attempts to measure Wikipedia's content-based "gender gap" have been focused on particular domains and do not generalize well across broad encyclopedic content. As mentioned above, Lam et al. [14] studied the quality of articles corresponding to movies rated highly by users who identify as either men or women, and other studies have focused on the number and breadth of women's biographies [e.g., 16,52,53]. While these studies provide insight and support the hypothesis of a kind of "gender gap" in content, they do not speak to a broad set of topics.

A "gender gap" in content is difficult to examine. To determine whether there is a "gender gap" in content, one must decide how content is—if it is—"gendered." Because the "gendering" of content is subjective and socially constructed, identifying a baseline set of content to test for gender affinity must pay careful attention to the conflation of biological sex and gender. However, from a pragmatic perspective of operationalization and of taking a "first cut" at this complex issue, a study needs to exploit perceptions of gender as a binary rather than as a continuum of a set of symbolic norms [19], or a performance [9]. Thus our goal in this case, was to use our method to assay whether Wikipedia reflects general representativeness of topics that are *presumed* to be of interest to an established readership of "women's" and "men's" periodicals.

Applying the Method

For our second case, we chose periodicals that had the largest distribution numbers in the U.S., were uniformly indexed by EBSCO, and were targeted to either men or women readers. (See Table 2.)

We then followed the same process described above for aggregating one year's worth (2014) of metadata for each periodical. We combined these lists of subject terms (or keywords) into a set for each category (e.g., "women's"). Initially, this gave us 4,567 terms from "women's" periodicals and 4,713 terms from "men's" periodicals. We then removed duplicates, leaving 4,039 unique terms from "women's" periodicals and 3,904 unique terms from "men's" periodicals. This gave us one list of subject terms (or keywords) for 2014 from the four "women's" periodicals listed above and a separate list for subject terms for 2014 from the four "men's" periodicals listed below.

Periodicals Marketed to Women	**Periodicals Marketed to Men**
Cosmopolitan	*Esquire*
Ladies Home Journal	*GQ*
Redbook	*Family Handyman*
Woman's Day	*Men's Health*

Table 2. Publications with the highest subscribed circulation in the U.S. by targeted readership (i.e., category).

We generated two sets of disjoint topic terms by removing all terms that were common to the set of terms (keywords) between the "men's" list and the "women's" lists.

We then used a random number generator to select 400 terms from each category list and manually searched for the resulting 800 terms using Wikipedia's native search feature. We recorded what we found using the heuristics described above for each search term and met to adjudicate imprecise matches.

Demonstrative Data

The results of our method for randomly selected terms from "men's" and "women's" periodicals are striking at a high level. We found that 67.6% of "women's" topics and 84.1% of "men's" topics were covered. This represents a 16.5% difference in the topical coverage of Wikipedia as it is represented from periodicals targeted to a specific "gendered" readership. Overall, 15.9% of "men's" topics had no corresponding article whereas 32.4% of "women's" topics had none. (See Figure 3.)

Considering the data a little more closely reveals some interesting details. First, our stance was to come up with heuristics that would select an article from Wikipedia if the article existed in some clear form. That is, we biased toward inclusivity. This is an interesting decision because the "direct match" heuristic yielded nearly the same number of articles for "men's" and "women's" presumed interests when taken in sum across both samples (159 articles of "women's" topics, 160 articles of "men's" topics). This illustrates there are a number of ways the issue of content representativeness can be operationalized—some of which will show a difference and some of which might not.

An alternative operationalization of representativeness might come from an exclusionary perspective that considers only when articles do not exist. That perspective tells a slightly different story than the inclusionary perspective with only 58 "men's" interest topics resulting in no possible match whereas 120 "women's" interest topics generated no possible match. From this perspective, there are twice as many missing articles of possible interest to readers of "women's" periodicals than there are missing articles of possible interest to readers of "men's" periodicals.

	Topics from Sources Marketed to Women		Topics from Sources Marketed to Men	
	count	%	count	%
Does not exist	120	32.43%	58	15.89%
Direct match	159	42.97%	160	43.84%
Direct hit edited term	1	0.27%	11	3.01%
Redirect	54	14.59%	85	23.29%
1st search result	18	4.86%	27	7.40%
2nd search result	8	2.16%	10	2.74%
Nth search result	6	1.62%	7	1.92%
Applied multiple rules	2	0.54%	3	0.82%
Match disambiguation	2	0.54%	4	1.10%
Compound keyword	2		3	
List	3		3	
Vague/Generic Term	18		26	
Duplicate/Variant	1		1	
Other disqualification	6		3	
Total terms	400		401	
Terms for analysis	370		365	
Term coverage		67.57%		84.11%

Figure 3. Summary of results for topics from sources marketed to women and men. The last 5 gray shaded rows were omitted from the analysis.

The "redirect" heuristic was the second most activated heuristic in this study. This presents yet another perspective on coverage. A redirect page is created when there is a belief that a term or phrase is directly synonymous with another, such that a user looking for the one topic most often means the other. There were 54 article pages identified from "women's" interest topics and 85 article pages identified from "men's" interest topics based on the redirect heuristic. This is an approximately 2:3 ratio and either represents that some content has not been carefully linked to a more likely topic of "women's" interests, or that the content is simply missing.

Lastly, the number of topics excluded from analysis (e.g., because they matched something that was not a legitimate article, because the topic could not be effectively disambiguated, or the topic was conceptually vague without access to the original material, etc.) was surprisingly stable at 30-35 topics per randomly sampled 400 topics.

Summary

Our preliminary findings reveal some gaps in the representativeness of content on Wikipedia corresponding to that *presumed* to be of interest to readers of periodicals targeted to women and men. We found some evidence that topics proven to be of commercial viability and notability and *perceived* to be of interest to readers of "women's" periodicals are not represented in Wikipedia. This work, of course, does not resolve the implications of the participation gap nor does it propose a solution to those problems in Wikipedia or other UGC systems. However, it does support the findings of existing work regarding gaps in specific kinds of content, and it suggests the English language Wikipedia does bias toward an inclusion of content *presumed* to be of interest to readers of "men's" periodicals.

DISCUSSION

Prior work that considers the content of UGC systems has often focused on understanding problems with content coverage, or a lack of coverage (a gap). Our contribution is a method that builds on the idea of coverage to consider topical coverage in the context of an identifiable potential readership or content consumer. We define that potential measure as *representativeness*. That is, we define representativeness as the proportion of content present in a UGC relative to the proportion of a potential consumer, audience or readership in a clearly identifiable population.

Our approach is designed to address methodological challenges present in some of the prior work. In particular, our method is designed to circumvent the problem of biases resulting from endogeneity. The approach relies on content sources external to a UGC that are deemed relevant as a function of sustained readership and circulation. While these external content sources may have their own editorial biases, the topical focuses of the sources provide a clear scope of relevant content. Further we mitigate potential editorial biases by relying on more than one exogenous source when collecting the relevant metadata.

As a side effect, our approach can be used to assay topical content coverage. Applying our approach without relying on readership populations or the relative readership demographic distributions can be interpreted as a measure of general topical coverage in a UGC. In this way our method is a more generalizable and repeatable method for judging topical coverage.

Relying on exogenous sources provides another benefit. Understanding the coverage of a UGC relative to some known external sources provides a way to inspect the potential content biases in both the external sources and the UGC. For example, in our demonstration cases, the editorial distinctions between the way proper names are handled in Wikipedia compared to standards established for the abstracting services would have allowed us to inspect some of the distinctions between well-known people and less well-known people. Specifically, in the case of well-known individuals it is likely that a Wikipedia editor has created a redirect page so that the "Lastname, Firstname" search works as well as the Wikipedia editorial standard of "Firstname Lastname."

Our method opens a number of possibly interesting research trajectories. In the way we executed the method, in both of our cases, we choose two points along a demographic continuum and implemented the method to create

distinctions between the populations. In the case of gender we selected "men" and "women" and for political ideology we selected "conservative" and "liberal." This particular choice is a convenience that makes the method a bit more manageable. However, we could have picked multiple, identifiable readership populations along those dimensions. The method would only have to be changed in one simple way. Instead of using the metadata and creating two disjoint sets of terms by removing the shared terms in the two metadata sets, we would create *n* disjoint sets of terms, one disjoint set for each identifiable population and its associated metadata sources. The disjoint sets could then be used following the rest of the described procedures.

Another potential research trajectory is to consider the shared interests of a given population. Our specific focus was to find representativeness that was a function of the, supposedly, special or exclusive interests of each given population. Another approach would be to understand the degree to which supposed, shared interests of groups within the population were present in a given UGC. The shared interests potentially identify the strengths of UGC and peer production communities. That is, the assumption is that UGC communities contribute content that is a function of the interests and expertise of the contributors. By focusing on the terms and content metadata at the intersection (instead of the disjoint content) we could understand more about the shared interests of the contributors and how that reflects the shared interests of the populations that might consume or read the content.

Yet another possible research trajectory would be to use our method to identify topics that have been almost exclusively targeted to one population (or readership) so that readers' actual rather than *presumed* interests in these topics can be interrogated through systematic sampling and repeated observations. For example, one might elect to ask readers—either directly or indirectly—questions about their political beliefs, and then measure the degree to which they are interested in topics *presumed* to be of interest to them when presented with the opportunities to read about these topics outside of the context of the original content sources.

Another extension of the method could consider frequencies of subject terms and the associated article readership in the UGC. That is, one could use additional metadata from the indexing database to generate the set of subject terms weighted by frequency of occurrence during a given year. In a UGC system like Wikipedia, page view data could be used to understand how frequently the topically associated page is requested. This data could then be used to understand whether editorially selected content from exogenous sources corresponds to the way users request given content from a UGC system.

There is an interesting issue here with the problem of judging frequencies as somehow equivalent to coverage density. It would probably be a mistake to take terms (or keywords) from an abstracting service and simply count their frequency in the targeted UGC system. For example, if the abstracting service was abstracting an article that talked about interesting new recipes that used apples, one of the terms generated might simply be "Apple (recipes)." Taking that and counting the frequency with which the term "apple" and "recipes" shows up in the UGC is probably not what one wants as a measure of topical density – because one has not measured a topic – one has measure a term frequency – and these are not the same.

One can also imagine modifying and applying our method to another UGC system like Quora, which has received media attention for having a both a known skew in participation and a misogynistic culture [5]. Instead of using a search feature, one could leverage Quora's lists of topical interests to investigate *representativeness* compared to exogenous sources, and then consider how sociotechnical features like up-voting and down-voting interact with topics *presumed* to be of more interest to women or men.

Finally, our method begins to expose technical and infrastructural aspects of a UGC system that may not be obvious to the average consumer of UGC system. One example has been briefly covered above, the nature of Wikipedia redirect pages. That a redirect handles differing rules of "Lastname, Firstname" or "Firstname Lastname" might not be terribly problematic. Some recent work has focused on the impacts of failing to account for redirects [24] in many of the quantitative studies of Wikipedia. However, our case raises another concern. In our method demonstration related to gender, the ratio of redirected content for "men's" versus "women's" topical terms yields a question about semantic control over a term and how that control can be made to bias for or against particular content.

Consider this claim a little further by recalling that all of the metadata we collected came from edited sources that had an article on the topic the term describes. This implies that all of the terms we were using had some reasonable claim to being a legitimate topic of an encyclopedic article. That there were more "men's" interest related redirects makes those topics more prominent and cedes control of the redirected term to one interest group to the exclusion of another. The alternative is that a term could generate a disambiguation page, which we also considered. The main point is that the method we describe can help raise consideration of these infrastructural aspects and make their implications more visible and, thus, more inspectable.

LIMITATIONS

While our general method avoids endogeneity bias and relies upon content proven to be commercially viable, we recognize an editorial team's choices to feature certain kinds of content is as subjective as those made by a group of contributors, or by an algorithm. We also recognize that all taxonomies—even those built upon principles of knowledge organization and library science—evidence some biases, but this is a larger issue beyond the scope of our study.

Admittedly our approach relies upon conceptualizations of political ideologies as simplistic, bounded, and mutually exclusive, and of gender as tied to the binary of man/woman, a view often perpetuated by mass media and popular culture. However, as mentioned above, the aim of the method is not to challenge definitions or binaries but rather to develop a systematic and reproducible approach to examine how content perceived as being of more interest to specifically targeted audiences is represented on influential UGC sites like Wikipedia. Further, we have mentioned above, a modification of the technique that would address interests at the intersection of one or more groups.

We also recognize some limitations of our specific implementation of the general method. We selected magazines with the highest distribution in English in the United States, and we sampled only one year (2014) worth of terms from the content keywords. Sampling over several years and including only topics that persist would most likely produce a more stable corpus for testing. Additionally, we tested our method using only the English language Wikipedia. Finally, our method relies upon human judgment rather than automation and is, therefore, challenging—but not impossible—to scale.

CONCLUSION

As a prominent and powerful UGC system, Wikipedia is a source of information propagated throughout numerous other sites and systems. Take, for example, Helix, a plug-in designed to help writers with research as they write by suggesting information from various online sources including Wikipedia [13]. Or consider that the ACM Digital Library now uses IBM Watson to pull "Concepts in this article" from Wikipedia's content. Consequently, the *representativeness* of Wikipedia's content has become increasingly important.

The significant contribution of this paper, then, is the description of an exogenous methodical approach to begin to unpack the representativeness of a UGC corpus by relying on an independently generated set of terms reflecting topical content aimed at specific demographics with proven commercial viability. This method generalizes to characterize *representativeness* of the content in a UGC system for other populations with which a researcher could associate identifiable, cataloged content providers. For example, one can see how this method could be used to identify representativeness of topical content aimed at children's interests, or those with interests in fitness and wellness, sports, religion, cooking, and others as these topical areas have multiple external content providers that are abstracted by content metadata service providers. More difficult to test, of course, is the kind of content socially marginalized in UGC systems like Wikipedia or by content providers in general, or content of interest to groups that are so small (relative to the population) that commercial providers do not exist.

Although challenging, developing—and sharing—a systematic method for measuring representativeness and, consequently, identifying potential gaps in content is an important task for researchers in CSCW and the broader HCI community. Despite the promises of increased access to and the normalization of social computing, we continue to see existing inequalities reproduced and even reified online. If we want to design systems that encourage people to engage in more pro-social or equitable ways and that serve as critiques of existing social norms, then we must first observe and understand existing behaviors and norms. Considering how UGC systems represent the presumed interests of different audiences and evidence gaps in content can help us examine potential relationships between diversity of participants and representativeness of content, implicit barriers to participation, and how popular, "open" UGC systems may reinforce—rather than challenge—exclusionary policies and practices.

ACKNOWLEDGMENTS

We would like to thank Brian Keegan for early feedback, and Siôn Romaine for his indispensible help with EBSCO. We would also like to acknowledge support from National Science Foundation (NSF) grant, IIS-1162114.

REFERENCES

1. Nate Anderson. (2007, March). Conservapedia hopes to "fix" Wikipedia's "liberal bias". *Ars Technica*. Retrieved from http://arstechnica.com/uncategorized/2007/03/conservapedia-hopes-to-fix-wikipedias-liberal-bias/
2. Annual estimates of the resident population for selected age groups by sex for the United States, states, counties, and Puerto Rico commonwealth and municipios: April 1, 2010 to July 1, 2014 Population Estimates. (n.d.). American FactFinder. Retrieved from http://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?src=CF
3. Judd Antin, Raymond Yee, Coye Cheshire, and Oded Nov. (2011). Gender differences in Wikipedia editing. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration* (WikiSym '11). ACM, New York, NY, USA, 11-14.
4. Patti Bao, Brent Hecht, Samuel Carton, Mahmood Quaderi, Michael Horn, and Darren Gergle. (2012). Omnipedia: bridging the Wikipedia language gap. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1075-1084.
5. Christie Barakat. (2014, May). Quora's misogyny problem. *Social Times*. Retrieved from http://www.adweek.com/socialtimes/quora-misogyny-problem/149508?red=st
6. Katherine Bischoping. (1993). Gender differences in conversation topics, 1922–1990. *Sex Roles 28*(1), 1-18.

7. Torie Bosch.. (2012, July). Kate Middleton's wedding gown and Wikipedia's gender gap. *Slate*. Retrieved from http://www.slate.com/blogs/future_tense/2012/07/13/kate_middleton_s_wedding_gown_and_wikipedia_s_gender_gap_.html
8. Adam R. Brown. (2011). Wikipedia as a data source for political scientists: Accuracy and completeness of coverage. *PS: Political Science & Politics 44*(02), 339-343.
9. Judith Butler. (1990). *Gender trouble: Feminism and the subversion of identity*. New York: Routledge.
10. Shou Chang, Vikas Kumar, Eric Gilbert, and Loren Terveen, L. (2014). Specialization, homophily, and gender in a social curation site: Findings from Pinterest. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing* (CSCW '14). ACM, New York, NY, USA, 674-686.
11. Sapna Cheryan, Jessica Schwartz Cameron, Zach Katagiri, and Benoît Monin. (2015). Manning up: Threatened men compensate by disavowing feminine preferences and embracing masculine attributes. *Social Psychology*.
12. Benjamin Collier and Julia Bear. (2012.) Conflict, criticism, or confidence: An empirical examination of the gender gap in Wikipedia contributions. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work* (CSCW '12). ACM, New York, NY, USA, 383-392.
13. Kate Conger. (2016, May). Helix conducts research as you write. *TechCrunch*. Retrieved from http://techcrunch.com/2016/05/08/helix-conducts-research-as-you-write/
14. Conservapedia: About. (n.d.). Retrieved from http://www.conservapedia.com/Conservapedia:About
15. Andrea Forte, Judd Antin, Shaowen Bardzell, Leigh Honeywell, John Riedl, and Sarah Stierch. (2012). Some of all human knowledge: Gender and participation in peer production. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work Companion* (CSCW '12). ACM, New York, NY, USA, 33-36.
16. Eduardo Graells-Garrido, Mounia Lalmas, and Filippo Menczer. (2015). First women, second sex: Gender bias in Wikipedia. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media* (HT '15). ACM, New York, NY, USA, 165-174.
17. Shane Greenstein and Feng Zhu. (2012). Is Wikipedia biased?. *The American Economic Review 102*(3), 343-348.
18. Alexander Halavais and Derek Lackaff. (2008). An analysis of topical coverage of Wikipedia. *JCMC 13*, 429-440.
19. Susan G. Harding. (1986). *The science question in feminism*. Cornell University Press.
20. Libby Hemphill and Jahna Otterbacher. (2012). Learning the lingo?: Gender, prestige and linguistic adaptation in review communities. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work* (CSCW '12). ACM, New York, NY, USA, 305-314.
21. Libby Hemphill, Jahna Otterbacher, and Matthew Shapiro. (2013). What's Congress doing on Twitter?. In *Proceedings of the ACM 2013 Conference on Computer Supported Cooperative Work* (CSCW '13). ACM, New York, NY, USA, 877-886.
22. Libby Hemphill and Andrew J. Roback. (2014). Tweet acts: How constituents lobby Congress via Twitter. In *Proceedings of the ACM 2014 Conference on Computer Supported Cooperative Work* (CSCW '14). ACM, New York, NY, USA, 1200-1210.
23. Benjamin Mako Hill and Aaron Shaw. (2013). The Wikipedia gender gap revisited: Characterizing survey response bias with propensity score estimation. *PloS one 8*(6), e65782. Chicago.
24. Benjamin Mako Hill and Aaron Shaw. (2014). Consider the redirect: A missing dimension of Wikipedia research. In *Proceedings of The International Symposium on Open Collaboration*. ACM, New York, NY, USA, 28.
25. Todd Holloway,, Miran Bozicevic, and Katy Börner. (2005). Analyzing and visualizing the semantic coverage of Wikipedia and its authors. ArXiv Computer Science e-prints, cs/0512085.
26. Kim Holmberg and Iina Hellsten.. (2014). Analyzing the climate change debate on Twitter: Content and differences between genders. In *Proceedings of the 2014 ACM Conference on Web Science* (WebSci '14). ACM, New York, NY, USA, 287-288.
27. Joshua L. Kalla and Peter M. Aronow. (2015). Editorial bias in crowd-sourced political information. *PloS one 10*(9), e0136327.
28. Matthew Kay, Cynthia Matuszek, and Sean A. Munson. (2015). Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (CHI '15). ACM, New York, NY, USA, 3819-3828.
29. Aniket Kittur, Ed H. Chi, and Bongwon Suh. (2009). What's in Wikipedia?: Mapping topics and conflict using socially annotated category structure. In *Proceedings of the SIGCHI Conference on Human*

Factors in Computing Systems (CHI '09). ACM, New York, NY, USA, 1509-1512.

30. Max Klein. (2015). Wikipedia in the world of global gender inequality indices: What the biography gender gap is measuring. In *Proceedings of the 11th International Symposium on Open Collaboration*. ACM, New York, NY, USA, 16.
31. Jona Kräenbring, Tika Monzon Penza, Joanna Gutmann, Susanne Muehlich, Oliver Zolk, Leszek Wojnowski, Renke Maas, Stefan Engelhardt, and Antonio Sarikas. (2014). Accuracy and completeness of drug information in Wikipedia: A comparison with standard textbooks of pharmacology. *PloS one 9*(9), e106930.
32. Shyong (Tony) K. Lam, Anuradha Uduwage, Zhenhua Dong, Shilad Sen, David R. Musicant, Loren Terveen, and John Riedl. (2011). WP:Clubhouse?: An exploration of Wikipedia's gender imbalance. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration* (WikiSym '11). ACM, New York, NY, USA, 1-10.
33. David Laniado, Andreas Kaltenbrunner, Carlos Castillo, and Mayo Fuster Morell. (2012). Emotions and dialogue in a peer-production community: The case of Wikipedia. In *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration* (WikiSym '12). ACM, New York, NY, USA.
34. Issie Lapowsky. (2016, May). Of course Facebook is bias: That's how tech works today. *Wired*. Retrieved http://www.wired.com/2016/05/course-facebook-biased-thats-tech-works-today/
35. Mike Ma. (2016, January). Wikipedia edit war over Google-featured activist's praise for Bin Laden. *Breitbart*. Retrieved from http://www.breitbart.com/tech/2016/05/20/wikipedia-editors-scrub-references-activists-bin-laden-praise-following-breitbart-article/
36. J. Nathan Matias, Sophie Diehl, and Ethan Zuckerman. (2015). Passing on: Reader-sourcing gender diversity in Wikipedia. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1073-1078.
37. Cynthia McKelvey. (2016, February). Political activists claim Twitter censored #WhichHillary hashtag. *The Daily Dot*. Retrieved from http://www.dailydot.com/politics/which-hillary-hashtag-twitter-censorship/
38. Sudip Mittal, Neha Gupta, Prateek Dewan, and Ponnurangam Kumaraguru. (2014). Pinned it!: A large scale study of the Pinterest network. In *Proceedings of the 1st IKDD Conference on Data Sciences* (CoDS '14). ACM, New York, NY, USA.
39. Amanda Menking and Ingrid Erickson. (2015). The heart work of Wikipedia: Gendered, emotional labor in the world's largest online encyclopedia. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (CHI '15). ACM, New York, NY, USA, 207-210.
40. Henry T. Moore. (1922). Further data concerning sex differences. *Journal of Abnormal Psychology 17*, 210-214.
41. Emma Pierson. (2015). Outnumbered but well-spoken: Female commenters in the New York Times. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (CSCW '15). ACM, New York, NY, USA, 1201-1213.
42. Ioannis Protonotarios, Vasiliki Sarimpei, and Jahna Otterbacher. (2016). Similar gaps, different origins?: Women readers and editors at Greek Wikipedia. In *Tenth International AAAI Conference on Web and Social Media*.
43. Joseph Reagle and Lauren Rhue. (2011). Gender bias in Wikipedia and Britannica. *International Journal of Communication 5*, 21.
44. Research gender affinity for different subjects on Wikipedia. (n.d.). Retrieved from https://meta.wikimedia.org/wiki/Grants:IdeaLab/Research_gender_affinity_for_different_subjects_on_Wikipedia
45. Lydia Saad. (2015, January). U.S. Liberals at record 24%, but still trail Conservatives. Gallup. Retrieved from http://www.gallup.com/poll/180452/liberals-record-trail-conservatives.aspx
46. Monika Sengul-Jones (2016). Final report. Retrieved from https://meta.wikimedia.org/wiki/Grants:IEG/Full_Circle_Gap_Protocol:_Addressing_the_'Unknown_Unknowns'/Final
47. Monika Sengul-Jones. (2016). The "bestiary of gaps" on Wikipedia. Gap finding project. Retrieved from https://gapfindingproject.wordpress.com/2015/10/12/the-bestiary-of-gaps-on-wikipedia/
48. Matthew Sheffield. (2008, August). Conservatives miss Wikipedia's threat. *The Washington Times*. Retrieved from http://www.washingtontimes.com/news/2008/aug/21/conservatives-miss-wikipedias-threat/?page=all
49. Eran Shor, Arnout van de Rijt, Alex Miltsov, Vivek Kulkarni, and Steven Skiena. (2015). A Paper ceiling: Explaining the persistent underrepresentation of women in printed news. *American Sociological Review 80*(5), 960-984.
50. Survey potential female editors to determine most popular topics of interest. (n.d.). Retrieved from https://meta.wikimedia.org/wiki/Grants:IdeaLab/Surve

y_potential_female_editors_to_determine_most_popular_topics_of_interest

51. Maja Van der Velden. (2013). Decentering design: Wikipedia and indigenous knowledge. *International Journal of Human-Computer Interaction 29*:4.
52. Claudia Wagner, David Garcia, Mohsen Jadidi, and Markus Strohmaier. (2015). It's a man's Wikipedia?: Assessing gender inequality in an online encyclopedia. In *Proceedings of the International AAAI Conference on Web and Social Media* (ICWSM2015).
53. Claudia Wagner, Eduardo Graells-Garrido, David Garcia, and Filippo Menczer. (2016). Women through the glass-ceiling: Gender asymmetries in Wikipedia. arXiv preprint arXiv:1601.04890.
54. Yi-Chia Wang, Moira Burke, and Robert E. Kraut. (2013). Gender, topic, and audience response: An analysis of user generated content on Facebook. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '13). ACM, New York, NY, USA, 31-34.
55. Wikipedia: Notability. (n.d.). Retrieved from https://en.wikipedia.org/wiki/Wikipedia:Notability
56. Wikipedia: Notability (people). (n.d.). Retrieved from https://en.wikipedia.org/wiki/Wikipedia:Notability_(people)
57. Jake Ryan Williams, Eric M. Clark, James P. Bagrow, Christopher M. Danforth, and Peter Sheridan Dodds. (2015). Identifying missing dictionary entries with frequency-conserving context models. *Physical Review E 92*(4), 042808.
58. Samuel Wooley and Phil Howard. (2016, May). Bots unite to automate the presidential election. *Wired*. Retrieved from http://www.wired.com/2016/05/twitterbots-2/