Measuring Self-Focus Bias in Community-Maintained Knowledge Repositories

Brent Hecht[†] and Darren Gergle[†‡]
[†]Dept. of Electrical Engineering and Computer Science
[‡]Dept. of Communication Studies
Northwestern University

brent@u.northwestern.edu, dgergle@northwestern.edu

ABSTRACT

Self-focus is a novel way of understanding a type of bias in community-maintained Web 2.0 graph structures. It goes beyond previous measures of topical coverage bias by encapsulating both node- and edge-hosted biases in a single holistic measure of an entire community-maintained graph. We outline two methods to quantify self-focus, one of which is very computationally inexpensive, and present empirical evidence for the existence of self-focus using a "hyperlingual" approach that examines 15 different language editions of Wikipedia. We suggest applications of our methods and discuss the risks of ignoring self-focus bias in technological applications.

Categories and Subject Descriptors

H.5.3 [**Information Systems**]: Group and Organization Interfaces – *collaborative computing, computer-supported cooperative work, theory and models.*

General Terms

Measurement, Experimentation, Human Factors, Languages

Keywords

Self-focus, bias, topical coverage, Wikipedia, Web 2.0, semantic networks, hyperlingual

1. INTRODUCTION

Allegations of bias have dogged user-contributed knowledge repositories, particularly Wikipedia, since the Web 2.0 revolution took off. These accusations have come from both popular media and academia. Stephen Colbert, in *The Colbert Report*, presented a playful description of "wikiality"—a form of bias in which users redefine the world according to their personal views. He urged his television audience to transform the Wikipedia entry on the "Elephant" to suggest that the population of elephants in Africa tripled over the last six months. This would then serve as a rather "inconvenient tusk" for Al Gore and his colleagues in the environmental movement. While this sketch was clearly intended as a comedy piece, it is illustrative of the central concern explored in this paper: Does an inherent bias based on shared personal opinions exist in community-maintained knowledge repositories?

C&T'09, June 25–27, 2009, University Park, Pennsylvania, USA.

ACM 978-1-60558-601-4/09/06.

In this paper, we explore this question by introducing and measuring *self-focus*, a new way of understanding a type of bias in the graph structures that underlie many of these community-maintained repositories, including one of the largest: Wikipedia. We define self-focus bias as occurring when contributors to a knowledge repository encode information that is important and correct to them and a large proportion of contributors to the same repository, but not important and correct to contributors of similar repositories.

Self-focus bias is similar to topical coverage biases [8, 11] in that it seeks to describe the semantic makeup of knowledge repositories. Topical coverage bias studies explicitly or implicitly compare the distribution of articles (or a similar measure) in particular semantic categories in Wikipedia to that of a more traditional knowledge repository, generally in an effort to show that Wikipedia describes in more detail semantic areas that are of interest to its users. However, self-focus goes beyond the traditional limits of topical coverage bias by evaluating not only what types of nodes exist in the graph (in Wikipedia, nodes are articles), but, critically, examining how these nodes exist in *relation* to other nodes via the graph's link structure. Said differently, in our work we take into account not simply whether an article exists, but also the prominence of that particular article within the network. In this way, self-focus is a broad-stroke holistic measure of the effect of shared opinions and interests in community-maintained Web 2.0 graph structures.

This work is also unique in that it uses an innately "hyperlingual" approach: we appraise and compare the self-focus that exists in 15 different language editions of Wikipedia, analyzing over eight million articles and 230 million links. No general coverage bias survey of so many of the leading Wikipedia editions has been undertaken.

We label this work "hyperlingual" because its findings are verified in not just one or a few languages, but in an ample selection of tongues from all over the world. In this way, the *validity* of the research is extended to a wide variety of languages and cultures, rather than being limited to fluent speakers of English. This is not the case in the vast majority of Wikipedia-related work. Notable exceptions include the research of Capocci et al. [4] and Ortega et al. [16].

This research contributes to both our theoretical and applied understanding. At a theoretical level, the work presents a formal description of self-focus bias, demonstrates its existence, and provides a computational method for measuring it. At an applied level, we note how self-focus bias has a number of implications for the development of technologies that rely upon community-maintained repositories. Without an understanding of the biases present in these knowledge repositories, we run the risk of developing technologies that fail. For instance, this work suggests that all technologies developed on the English Wikipedia (and evaluated on English data sets) need to be adapted to or re-run on other language editions of Wikipedia in order to function well in the context of other languages and cultures.

In the following pages, we describe this paper's standing in the context of Wikipedia bias research. We then explain the implementation of the base system for our studies. We present two general methods used to identify and measure self-focus across the 15 Wikipedias, followed by a demonstration of the methods and strong *empirical evidence* for the existence of self-focus in Wikipedia. After establishing this, we apply these methods to a test case in the political domain to illustrate the broader impact self-focus studies can have. Finally, we discuss several issues brought to light during the analysis stage and present future research avenues.

2. BACKGROUND LITERATURE

Denning and colleagues developed a general framework for thinking about the biases present in Wikipedia [5]. In their work, they suggest six classes of risks: accuracy, motives, uncertain expertise, volatility, coverage, and sources. They write that these risks are taken on by any user of Wikipedia that assumes they are accessing the "entire range of [truthful] human knowledge, past and present." In other words, their risks can be assumed to be an enumeration of sources of *bias*, at least in the common parlance.

In the context of Denning et al.'s work, self-focus can be considered a seventh Wikipedia bias. As noted above, while a unique and powerful shaping factor in and of itself, self-focus also plays an important role in four of Denning et al.'s other bias sources – accuracy, uncertain expertise, motives, and, particularly, coverage – just as Denning et al.'s sources have interplay amongst themselves.

To our knowledge, this research is not only the first to identify and study self-focus, but it also represents a more automated and flexible Wikipedia bias test than many in the literature [7, 8]. More importantly, however, this work is the first to measure bias by comparing language editions of Wikipedia *with each other*. In total, 15 different language editions of Wikipedia – hereafter referred to as individual *Wikipedias* – are examined. However, the method is expandable to any language edition of Wikipedia. Solely utilizing each Wikipedia's innate graph structure (rather than relying upon language-dependent features) accomplishes this flexibility.

In recent years, there have been a number of papers in addition to Denning et al.'s that have studied various aspects of Wikipedia bias. Famously, Giles found that Wikipedia stacked up relatively well against *The Encyclopedia Britannica* in the category of scientific articles [7]. Halavais and Lackaff studied Denning et al.'s topical coverage variable [8], noting, along with Holloway and colleagues [11], that topics of interest of Wikipedia's authors tend to be better covered in terms of number of articles. However, they stop short of analyzing the full network systemic effect that we have labeled self-focus. On the other hand, Bellomi and Bonato perform basic network analyses on an early version of the English Wikipedia [1] in an attempt to also uncover "hidden" biases, but do not consider the semantic node distribution in their study. In addition, none of these studies analyze anything but the English Wikipedia.

3. DATA PRE-PROCESSING

For the first stage of data pre-processing, we follow techniques described in previous work [9, 10] for analyzing a single Wikipedia. However, additional pre-processing is necessary when alignment of multiple Wikipedias needs to occur. Figure 1 explains the final alignment of the data, and the process is detailed below.

Table 1. Basic statistics of the numbers of included nodes and edges in each of the fifteen Wikipedia Article Graphs in this research.

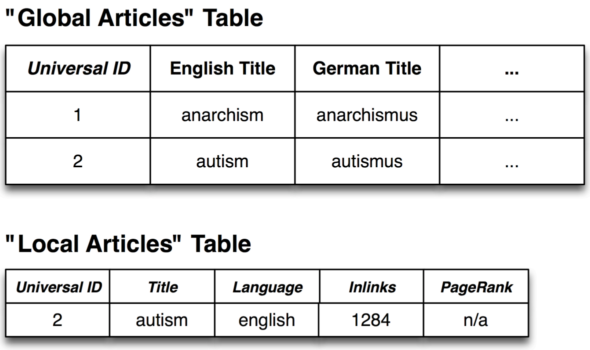
Language	Nodes (Articles)	Edges (Links)
Catalan	142,361	2,828,864
German	904,876	20,440,958
English	2,568,133	77,159,784
Spanish	449,816	10,420,933
Finnish	203,869	3,387,014
French	789,252	19,856,057
Italian	557,479	13,108,517
Japanese	553,578	21,187,662
Dutch	543,013	8,637,714
Norwegian	204,158	3,337,234
Polish	578,545	11,894,531
Portuguese	476,597	7,710,601
Russian	369,707	6,350,320
Swedish	310,048	4,974,216
Chinese	225,089	5,575,674
TOTAL	8,876,521	216,760,079

The root of the alignment process is comprised of the "interlanguage links" (ILLs) that exist in all Wikipedias. These take the form of "[[lang_code: title of article in the Wikipedia of lang_code]]". For instance, an interlanguage link that is placed in the English article "Pennsylvania State University" targeted at the French Wikipedia, whose language code is "fr", is "[[fr:*Université d'État de Pennsylvanie*]]". While there are automated bots that attempt to generate the transitive closure of these interlanguage links across the different Wikipedias, these bots are not 100 percent effective. In addition, our source data – the database dumps provided by the Wikimedia Foundation[1] – are generated at different times for each Wikipedia, leading to greater inconsistencies between the ILLs.

As such, article alignment is not a straightforward process. A naïve greedy algorithm that simply aligns articles based on the first interlanguage link found results in a large number of errors of omission and commission. We developed a more advanced, two-pass algorithm to reduce the number of errors. The first pass is similar to the initial greedy algorithm, but uses a small number of input languages as "ground truths" (we used English) in order to catch ILLs not propagated by the bots. The second pass checks to see if the initial greedy storage of interlanguage links missed any articles or referenced any that have since been renamed or deleted.

[1] All data dumps used in this study come from the fall of 2008.

It is important to note that it is possible for multiple articles in each language to have more than one article per universal identification number (see figure 1). We do not consider this a bug, because, for example, a single entry in English, might be better split into two entries in Japanese.



"Global Articles" Table

Universal ID	English Title	German Title	...
1	anarchism	anarchismus	...
2	autism	autismus	...

"Local Articles" Table

Universal ID	*Title*	*Language*	*Inlinks*	*PageRank*
2	autism	english	1284	n/a

Figure 1. A demonstration of the structure of the database that contains the aligned multiple language editions of Wikipedia.

4. MEASURES OF SELF-FOCUS

We introduce two different methods to identify and measure self-focus, both based on the Wikipedia Article Graph (WAG), or the web of links, of each Wikipedia. A key element of both methods is article *indegree*, or number of inlinks per article. Inlinks have a more salient meaning in Wikipedia than they do in the Internet as whole: a link from one article in Wikipedia to another is representative of a relation between the concept being written about and the concept being linked to. When a contributor to the French Wikipedia links the article "*Seconde Guerre Mondiale*" (World War II) to *"*Juno Beach*"*, s/he is encoding a relationship from World War II to Juno Beach in the French Wikipedia. If this link is not present – or if the article on Juno Beach does not exist – say, in the Catalan Wikipedia, no explicitly coded relationship between World War II and Juno Beach is evident to the Catalan Wikipedia reader (or to any automated process analyzing the Catalan Wikipedia). Here, we begin to see both the node- and edge-based components of self-focus.

It is a primary assumption of both methods that more inlinks to an article indicates more encoded relationships to that article, which therefore demonstrates a greater *focus* of a Wikipedia's WAG on that article. *In other words, an article with a lot of focus is an article that a Wikipedia's contributors have concluded to be very related to the sum of world knowledge represented in the other articles of the Wikipedia*. In this paper, we will show that this focus tends to be at least partially comprised of *self-focus*: people encoding relationships that are perhaps important and correct to them in their internal knowledge representation, but not important and correct to contributors to other Wikipedias. When enough people editing the same Wikipedia exhibit the same patterns, this amounts to an enormous bias within that Wikipedia when *compared to other Wikipedias* (the "similar repositories" in the terminology of the definition of self-focus found in the introduction).

Certainly, there are a large number of relationships that are universally important to contributors across all the Wikipedias. One example is the relationship between "World War II" and "United States". However, we show that these universally encoded relationships form a smaller proportion of Wikipedia links than one might have previously anticipated.

It is important reiterate that a link – an encoded relationship – must come from somewhere. For instance, the link to "Juno Beach" from "*Seconde Guerre Mondiale*" exists solely because there is an article on World War II in the French Wikipedia. Had there not been such an article, or if that article did not discuss D-Day battle locations, no link would exist. As such, an article (node) gains a lot of focus if *both* relevant articles link to the article *and* a sufficient number of those relevant articles exist. A single article's focus, and thus its potential self-focus, is thereby dependent on all aspects of the WAG: nodes and edges.

The first method we use to measure self-focus, that of *indegree summation*, is a simple-but-powerful measure that involves adding up the inlinks to specific groups of articles and comparing that sum across different language editions of Wikipedias. The second method, *PageRank score summation*, uses the PageRank algorithm [2] in the same way. In PageRank, articles with greater inlinks are given greater weight, so this can be considered a *weighted* form of indegree summation, with more general articles given more sway. For instance, in the English Wikipedia, links originating in the "Barack Obama" article would be weighted more heavily than links originating in "Chicken, Alaska".

5. STUDY 1: SELF-FOCUS AS INDEGREE SUMS

In this study, we determine whether in each language edition of Wikipedia a certain part of the world has greater *focus*. If that part of the world comprises the home region of each Wikipedia's language in a large number of Wikipedias, then empirical support for self-focus has been provided. In this case, we measure focus by the number of inlinks directed at Wikipedia articles located in a particular region of the world via indegree summation. We are able to determine the location of Wikipedia articles (if they have a location) via the latitude and longitude tags included in *spatial articles* by Wikipedia users[2]. Since Wikipedia's goal is to encode all of world knowledge, a greater focus, measured as such, would indicate an innate belief in a particular Wikipedia that the part of the world in greater focus is more *related* to the sum of all world knowledge than other places. Our hypothesis is that the "home region" of each Wikipedia – that region where the language of the Wikipedia is either primary and/or has a significant number of speakers – will be the geographic focus of each Wikipedia to a large extent. In other words, we predict that each Wikipedia will exhibit a large degree of *self-focus*. This is the opposite of the hypothesis of global consensus: that all Wikipedias will agree about which parts of the world deserve the most focus.

We first mapped all spatial articles that exist in each Wikipedia. We then performed a spatial join – an operation from the Geographic Information Systems (GIS) domain – to sum the indegrees of all articles in a given region. For instance, if the United States contained only two spatial articles in the Chinese Wikipedia (it, of course, actually contains many more), each with an indegree of 7, the United States would be assigned an *indegree*

[2] We used the data set from the WikiProject "Georeferenzierung".

sum of 14. We performed this analysis at two spatial scales: that of the country and that of the first-order administrative district (i.e. state, province, etc.). In other words, we calculated the indegree sums over both the set of countries and the set of administrative districts[3].

5.1 Results

The initial result from our study is that the indegree sums for the different Wikipedias at both the country and administrative district scales have an alarmingly low level of correlation with each other for data sets that are all supposedly trying to encode the exact same mass of world knowledge. In other words, there is very little correlation in focus on the country or administrative district scales across the Wikipedias. The average correlation of indegree sums at the country level between the Wikipedias was 0.365. At the administrative district level, it was at about the same low level (0.348). In addition, correlation with population values is also very low; in other words, indegree sums are *not proportional to the population of the country or administrative district*. The average correlation between indegree sum and population across the Wikipedias was 0.183 at the country level and 0.134 at the administrative district level. The only relatively high correlation with population occurred with the Chinese Wikipedia (0.65), which can be chalked up to a coincidence of self-focus and population (this correlation drops to 0.356 in administrative district analysis). Without Chinese, the average correlations with population were 0.151 and 0.120, for countries and administrative districts respectively. Obviously, other factors are entering into Wikipedians' linking strategies than previous practice of other Wikipedias and general interest measures like population.

However, another interesting pattern we noted in the indegree summation correlation tables was that languages with high cross-language fluency have higher correlations in indegree sums. For example, Catalan and Spanish are the most correlated of any two Wikipedias (0.950 country, 0.658 administrative district). English and Swedish (0.788, 0.733), and Portuguese and Italian (0.725, 0.764) are also highly correlated. One possible explanation is that a large amount of translation, instead of independent knowledge generation, is taking place between these Wikipedias. When people can translate existing articles with a high degree of fluency, they bring with the translation the knowledge representation from the original language. Of course, an element of shared self-focus is likely behind this phenomenon as well, as people who speak the same language are able to communicate culture much more effectively (and may even inhabit the same space, as is the case with Spanish and Catalan).

Now that a lack of *agreement* in focus has been established, it is possible to examine whether the difference in focus is due in part to self-focus. To do so, we introduce the simple *self-focus ratio* (SFR), which, if above 1.0, is a rather definite guarantee of a self-focus effect. It was calculated as follows:

$$SFR(W_{L=l}) = \frac{\max(C_{L=l})}{\max(C_{L\neq l})}$$

where C is the indegree sum of a country (or other region), $W_{L=l}$ is the Wikipedia of language l, and $C_{L=l}$ is the indegree sum of a country (or other region) where l is widely spoken. In other words, the self-focus ratio is the ratio of the maximum indegree sum of a country with language l to the maximum indegree sum of a country whose predominant language is *not* that language.

Table 2. The countries with the top indegree sums in the English (left) and French (right) Wikipedias.

Country	Indegree Sum
United States[4]	1,366,261
United Kingdom	439,582
France	189,698
Germany	151,303
Canada	146,191
Italy	129,133
Australia	127,539

Country	Indegree Sum
France	489,999
Italy	116,544
United States	80,876
Switzerland	49,999
Germany	48,828
Spain	44,611
United Kingdom	39,158

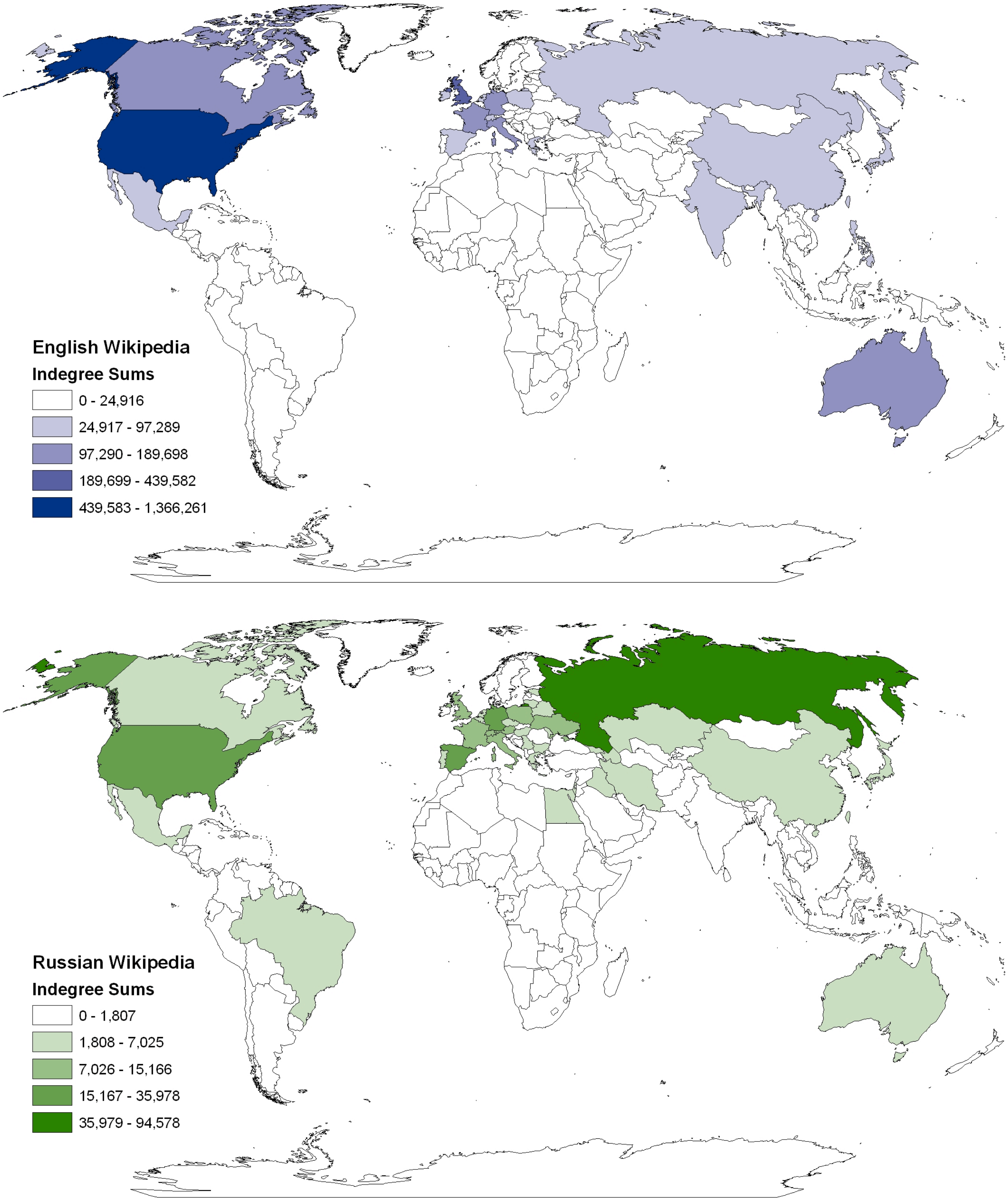
Let us consider the English Wikipedia as an example (L = English). The country with the greatest indegree sum of the countries where English is widely spoken is the United States, with an indegree sum = 1,366,261 = $\max(C_{L=l})$. As can be seen in table 2, the second greatest indegree sum belonged to the United Kingdom. France, which obviously does not use English as a predominant language, had the third greatest indegree sum = 189,698 = $\max(C_{L\neq l})$. The rest is a simple ratio calculation. To give the reader a better idea of the numbers behind the SFRs, tables 3 and 4 list the countries with the top indegree sums for the Japanese and Finnish Wikipedias. The spatial distribution of indegree sums in the English and Russian Wikipedias is shown in maps 1 and 2. Across all of these data, it is readily apparent that the Wikipedia for a particular language shows a distinct pattern in which the countries where that language is popular are the targets of strong self-focus bias.

As table 5 shows, our hypothesis of self-focus proved very correct in twelve of the fifteen Wikipedias, and correct but to a less extreme degree in the other three. The twelve are discussed first, followed by an analysis of the final three.

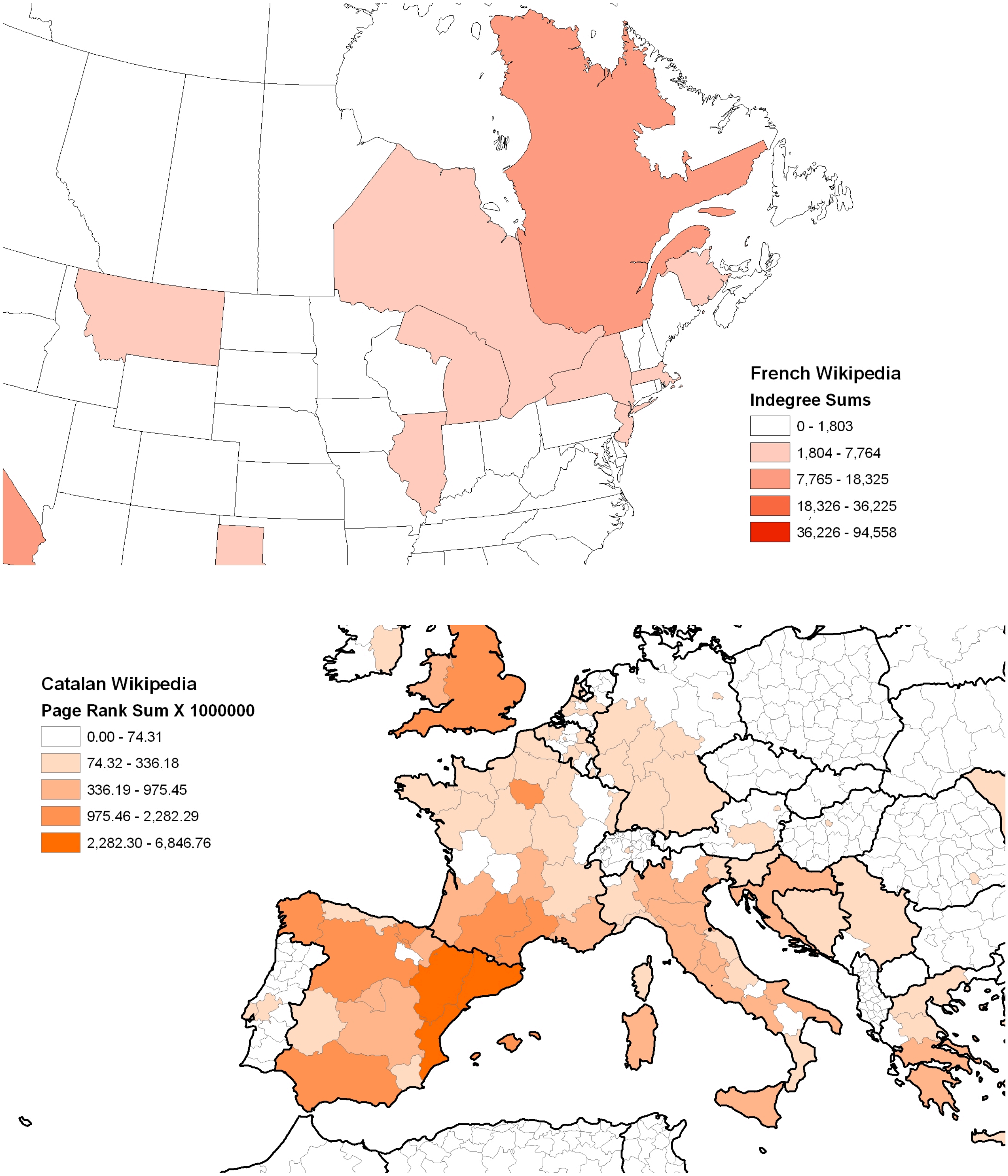
In short, a Wikipedia with an SFR above 1.0 is a Wikipedia that represents the sum of its world knowledge as being focused more on a home region country than on any other country in the entire world. This is extreme self-focus. Of particular interest are languages with relatively tiny numbers of speakers such as Finnish, in whose Wikipedia the country of Finland, and its only 5 million residents, is easily the predominant spatial focus of the entire Wikipedia. The nearest competitor to Finland in the

[3] We did not average these sums over number of articles because doing so would assume a specific configuration of self-focus: a small number of highly-inlinked articles. This is, in fact, not the predominant expression of self-focus.

[4] As a side note, there is a chance that for at least the English Wikipedia, the indegree sum is slightly inflated, because a bot early on in Wikipedia's history added spatial containment relations to many articles on United States cities and towns. (For instance, it added a link from Chicago to Cook County). It is impossible to know how many of these links would exist normally, but likely a high proportion.



Maps 1 and 2. Indegree sums of the English Wikipedia (top) and the Russian Wikipedia (bottom), calculated at the country-level. Because classes were determined via Jenks' Natural Breaks algorithm, the extreme outlier nature of the United States and Russia can be seen clearly.



Maps 3 and 4. Indegree sums of the French Wikipedia (top) and PageRank score sums of the Catalan Wikipedia (bottom), calculated at the administrative district-level. Because classes were determined via Jenks' Natural Breaks algorithm, the extreme outlier natures of Quebec and Cataluña (and its neighbors) can be seen easily.

Finnish Wikipedia is the United States and its over 300,000,000 residents, which has less than half the inlinks directed at it as Finland. China is not even in the top five. Considering an example from one of the larger Wikipedias, Japan has approximately *6.4 times more links directed at it* than the second-place country in inlinks summation in the Japanese Wikipedia, Italy.

Table 3 and 4. The countries with the top indegree sums in the Japanese Wikipedia (top) and the Finnish Wikipedia (bottom).

Country	Indegree Sum
Japan	453,048
Italy	70,922
United States	60,384
China	37,208
Germany	25,276

Country	Indegree Sum
Finland	55,331
United States	25,664
Germany	11,972
Russia	10,076
United Kingdom	9,402

Table 5. The *self-focus ratio* of each Wikipedia, as described above.

Language	Self-Focus Ratio
English	7.2
Japanese	6.4
German	6.3
French	4.2
Italian	3.6
Catalan	2.9
Russian	2.6
Spanish	2.4
Finnish	2.2
Polish	1.7
Norwegian	1.4
Chinese	1.2
Dutch	0.7
Swedish	0.6
Portuguese	0.3

Where our hypothesis does not prove so obviously true, there may be extenuating circumstances. In the case of both the Dutch and the Swedish Wikipedias, the United States was the indegree sum leader. However, the home countries had the second greatest indegree sums in each case, *which still represents a very large amount of self-focus*. The lower SFR ratio could be explained in that the Dutch and Swedish societies are both highly bilingual with English and may have gained significantly more guidance from the English Wikipedia, muting their spatial self-focus effect. Similarly, they could simply be more interested and/or aware of locations within the United States, as well as topics that are related to these locations (thus increasing their focus). In the case of Portuguese, the countries with a higher indegree sum than Brazil are Italy and the United States, indicating a possible peculiarity with the contributions to the Portuguese Wikipedia (a bot, for example, may be the cause). Regardless, it is important to remember that Brazil is still the third-largest destination of links in the world, indicating a large amount of self-focus despite the smaller SFR.

Fewer decisive conclusions can be drawn from the administrative district-level analysis of spatial indegree sums. However, there is strong anecdotal evidence for our hypothesis at this scale. Consider maps 3 and 4, which show the notable self-focus on Quebec in the French Wikipedia and Catalonia and its neighbors in the Catalan Wikipedia.

6. STUDY 2: SELF-FOCUS AS PAGERANK SCORE SUMS

This experiment is quite similar to the previous one, except simple indegree sums have been exchanged for *PageRank score sums*. If all the self-focus in spatial indegree were coming from peripheral articles (like “Chicken, Alaska” rather than “Coca-Cola”), the results from this experiment would differ significantly from those in the previous section. However, if self-focus permeates the WAGs throughout, regardless of the importance of articles, similar results should be expected.

Although we were only able to run the PageRank algorithm on the three smallest Wikipedias in our study (Catalan, Finnish and Norwegian) due to the computational complexity of the PageRank algorithm and the extensive size of the large Wikipedias (see table 1), results from this experiment suggest the latter is true; we see similarly strong self-focus patterns with PageRank sums as we did with indegree sums. Of course, further research is needed to definitively resolve this question. The analogue of table 5 is found in table 6, and a map showing an illustrative example is shown in map 4.

Table 6. The PageRank sum *self-focus ratio* of each Wikipedia, as described above.

Language	SFR with PageRank
Catalan	2.7
Finnish	1.7
Norwegian	0.5

7. APPLICATIONS OF SELF-FOCUS MEASUREMENT

With indegree sums and PageRank sums shown to be a satisfactory indicator of self-focus, it is now possible to look at these measures as windows into the predominant internal interests and opinions of Wikipedia contributors. While these measures may be surprisingly simple, it is demonstrated above that it is reasonable to infer self-focus from them. Of course, it would be impossible to confirm these opinions and interests without a massive survey of all Wikipedia contributors, but it is educational to see what kind of relative biases indegree sums and PageRank sums suggest in important domains such as politics. This is a major direction of future research, but preliminary results are promising.

Table 7 shows the ratio of inlinks to the “Barack Obama” article to those to the “John McCain” article in a selection of the Wikipedias studied. Given that indegree sums proved to be such convincing proxies of self-focus in the previous section, one might also conclude that the vast majority of Wikipedias' contributors considered Barack Obama to be more important to the rest of world knowledge than John McCain, even though all of the Wikipedia database dumps were gathered prior to the completion of the United States presidential election of 2008.

Another domain we are exploring is that of the European Parliament. Initial results have shown that indegree sums to European Parliament parties differ in each Wikipedia, and may even differ extensively from the political distributions of each Wikipedia's home countries' delegations, suggesting yet more interesting uses of indegree and PageRank sums as proxies for self-focus bias.

Table 7. The ratio of inlinks to the article for Barack Obama to those to the article for John McCain, by Wikipedia

Language	Inlinks(Obama)/Inlinks(McCain)
German	1.27
English	1.31
Spanish	1.08
French	1.34
Italian	1.22
Japan	1.14
Dutch	1.31
Norwegian	2.23
Polish	0.71
Portuguese	2.00
Russian	0.74
Swedish	0.76
Chinese	1.91

8. DISCUSSION

How does self-focus permeate across the various Wikipedia Article Graphs (WAGs)? This is a key question raised by this research. In other words, while we have shown and measured the end effect of self-focus on the WAGs – and we assume the original cause to be differentials in collective opinions about what is interesting and correct between the Wikipedias – what are the intermediary processes that put this effect into place? Most critically, does the main difference in the WAGs reside in the links between articles that exist in all languages or links to articles that exist in fewer than the entire fifteen? Preliminary results indicate both are contributing factors. In an initial repeat of the spatial indegree sum study, country-scale correlations went up by a large margin when only articles that exist in all 15 languages were considered, but the correlations were still far from 1.0 (given the much smaller sample size, large outliers like the United States were removed). In other words, if the set of spatial articles were limited to those spatial articles that exist in all languages, contributors to different Wikipedias would link to these articles at different rates, but not at the massively differentiated rates one might expect from tables 5 and 6.

Why did the correlations go up by a large margin when we took out the spatial articles that did not exist in all 15 Wikipedias? The example of "Chicken, Alaska", which only exists in the English, French, Dutch, and Portuguese Wikipedias, is illustrative. Once this article is created in these languages, it is likely linked to the articles on nearby Alaska Route 5 and the Taylor Corridor (and these articles might be created if they do not already exist). Links to the articles on the United States and the state of Alaska are also more or less mandatory (a spatially-dependent version of preferential attachment [4]). These two processes create a disproportionate growth of indegree sums for both the state of Alaska and the United States in the Wikipedias in which the "Chicken, Alaska" article exists compared to those in which it does not (This will also occur with non-spatial articles that do not appear in all Wikipedias, although likely to a lesser extent). Although more research is needed, these findings indicate that the topical coverage biases in article count suggested in [8] are at least partially responsible for the creation of self-focus in the Wikipedia network.

The subject of Africa brings up an entirely different area of discussion. In *none* of the Wikipedias does any country in Sub-Saharan Africa contain significant indegree or PageRank sums (see maps 1 and 2). While this study has experimentally shown self-focus to be a powerful centralizer of indegree on the home region, the innate reverse effect of this phenomenon is a defocusing on other areas. Africa, as in so many other domains, gets the short end of the stick, likely due to both a dearth of links to articles that exist about Africa, as well as a limited number of such articles. Like it is in physical world, this study shows that Africa is unfortunately on the periphery of Wikipedia.

9. CONCLUSION AND FUTURE WORK

In the time since the publication of [5], Wikipedia has become far more than just an extraordinarily popular web site. It and other Web 2.0 resources are now the key data repositories in critical new systems such as [6, 9, 10, 13-15, 17, 18] and the source of new knowledge in human behavior and human-computer interaction [3, 4, 12]. While the risks identified by Denning et al. must be taken as assumptions by all of these inventions and discoveries, self-focus, too, must be considered to be a risk. Even though this study utilized the article network (WAG), other structures of Wikipedia – such as article word vectors and the category network (WCG) – are affected due to the fact that, as noted above, "links must come from somewhere". While future research will seek to evaluate these effects in greater detail, self-focus should be an important concern in any Wikipedia-based application and discovery, regardless of the structure of Wikipedia utilized. For instance, our study suggests that one of the most known applications of Wikipedia, the article word vector-based "semantic relatedness" measure developed by Gabrilovich and Markovitch [6] is very biased towards English, and, more importantly, the people who speak it well enough to contribute to the English Wikipedia.

That said, there is still much to learn about self-focus. First and foremost, it would be helpful to have a model that could relatively accurately predict the existence of self-focus from external variables. Our results suggest that population is not, for the most part, a causational factor for indegree and PageRank score summations of spatial units. Analogously, this likely means the global size of interest groups is also not a good predictor of self-focus. A multivariate model capable of predicting self-focus is needed.

Second, the diffusion process of self-focus from the Wikipedia contributor level, to the group process level, all the way through to the graph-wide effect must be fully explicated. Future work should weigh the various factors examined in section eight so as to provide another way to predict self-focus at a more graph-theoretical level.

As the theory and causes of self-focus begin to be fully explained, it will also be beneficial to analyze self-focus in Wikipedia more thoroughly in controversial domains such as politics. Similarly, tools could be constructed that would suggest work that can be done to better balance focus in these domains, giving back to Wikipedia along the lines of Weld and colleagues in [18].

Finally, we would like to again highlight the importance of the hyperlingual approach and the opportunities it provides. In the past few years, researchers have made important observations about the way humans write, represent knowledge, work together, and more, using data from Wikipedia. However, with a few rare exceptions, existing Wikipedia work is limited to a single language, almost always English. Using research software such as that developed for this work[5], in many cases it is nearly as easy to gather data from and draw conclusions about 15 (or more) Wikipedias as it is one Wikipedia. If a researcher's conclusions hold across all these Wikipedias, it creates a much stronger case for her/his results. Our forthcoming work includes a large number of hyperlingual projects, and we hope to gain more company in the future.

10. ACKNOWLEDGMENTS

This research was funded, in part, by National Science Foundation grant #0705901 and Microsoft Research. A particularly special and grand thank you to Nada Petrović for all of her in-depth and valuable comments. Thanks also go out to Brian Keegan and Emily Moxley for their wise input on this paper and to our reviewers for their valuable comments.

REFERENCES

[1] Bellomi, F. and Bonato, R. (2005). Network Analysis for Wikipedia. *The First International Wikimedia Conference (Wikimania '05)*.

[2] Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Seventh International World Wide Web Conference (WWW '98)*, 107-117. Elsevier B.V.

[3] Buriol, L.S., Castillo, C., Donato, D., Leonardi, S. and Millozzi, S. (2006). Temporal Analysis of the Wikigraph. *Web Intelligence (WI '06)*. Washington, DC, USA: IEEE Computer Society.

[4] Capocci, A., Servedio, V.D.P., Colaiori, F., Buriol, L.S., Donato, D., Leonardi, S. and Caldarelli, G. (2006) Preferential attachment in the growth of social networks: The internet encyclopedia Wikipedia. *Physical Review E*, *74* (036116).

[5] Denning, P., Horning, J., Parnas, D. and Weinstein, L. (2005) Wikipedia Risks. *Communciations of the ACM*, *48* (12). 152.

[6] Gabrilovich, E. and Markovitch, S. (2007). Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. *Twentieth Joint Conference for Artificial Intelligence (IJCAI '07)*, 1606-1611.

[7] Giles, J. (2005) Special Report: Internet encyclopaedias go head to head. *Nature*, *438*. 900-901.

[8] Halavais, A. and Lackaff, D. (2008) An Analysis of Topical Coverage of Wikipedia. *Journal of Computer-Mediated Communication*, *13* (2). 429-440.

[9] Hecht, B. and Raubal, M. (2008). GeoSR: Geographically explore semantic relations in world knowledge. *Eleventh AGILE International Conference on Geographic Information Science (AGILE '08)*, 95 - 114. Berlin, Germany: Springer-Verlag.

[10] Hecht, B., Starosielski, N. and Dara-Abrams, D. (2007). Generating Educational Tourism Narratives from Wikipedia. *Association for the Advancement of Artificial Intelligence Fall Symposium on Intelligent Narrative Technologies (AAAI-INT '07)*, 37-44.

[11] Holloway, T., Bozicevic, M. and Börner, K. (2007) Analyzing and visualizing the semantic coverage of Wikipedia and its authors. *Complexity*, *12* (3). 30 - 40.

[12] Kittur, A., Chi, E., Pendleton, B.A., Suh, B. and Mytkowicz, T. (2007). Power of the Few vs. Wisdom of the Crowd: Wikipedia and the Rise of the Bourgeoisie. *Computer-Human Interaction (CHI '07)*, 1-9.

[13] Mihalcea, R. (2007). Using Wikipedia for Automatic Word Sense Disambiguation. *NAACL-HLT*, 196-203.

[14] Milne, D. and Witten, I.H. (2008). An effective, low-cost measure of semantic relatedness obtained from Wikipedia. *AAAI 2008 Workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy (WIKI-AI '08)*.

[15] Nguyen, D.P.T., Matsuo, Y. and Ishizuka, M. (2007). Subtree Mining for Relation Extraction from Wikipedia. *NAACL-HLT*, 125-128.

[16] Ortega, F., Gonzalez-Barahona, J.M. and Robles, G. (2008). On The Inequality of Contributions to Wikipedia. *41st Annual Hawaii International Conference on System Sciences (HICSS '08)*, 304-311.

[17] Richman, A.E. and Schone, P. (2008). Mining Wiki Resources for Multilingual Named Entity Recognition. *Association for Computational Linguistics (ACL-08: HLT)*, 1-9.

[18] Weld, D.S., Wu, F., Adar, E., Amershi, S., Fogarty, J., Hoffman, R., Patel, K. and Skinner, M. (2008). Intelligence in Wikipedia. *Twenty-Third Conference on Artificial Intelligence (AAAI '08)*.

[5] Our Wikipedia API and analysis tool, "WikAPIdia" is available to researchers upon request. E-mail one of the authors for more information.