Global Attention Profiles – A working paper
First steps towards a quantitative approach to the study of media attention
Ethan Zuckerman

Summary
News media outlets (newspapers, radio and television broadcasts and websites) have finite capacities. Newspapers have practical limits to the number of articles that can be printed each day. Radio and television broadcasters can fit only so many stories into a 30 minute newscast, and news websites must select which stories fit on their homepages.

The genesis of this paper was the anecdotal observation that major English-language news media outlets devote more attention to some countries than to others. This is to be expected: in a given week, some countries will experience newsworthy events like wars, natural disasters, scientific discoveries, economic collapses, sports championships, while others will not. But it is equally clear, on an anecdotal basis, that some countries get far more attention on a consistent basis, without regard to the relative frequency or magnitude of newsworthy events.

How do newspapers, newscasts and website divide their attention between regions of the world? To which countries to they devote the most attention? Why do some countries get more attention than others? Do factors like a country's population and the size of its economy predict which countries will command the most attention from media channels?

This paper begins to answer some of these questions with repeatable, transparent statistical tools. It introduces the Global Attention Profile (GAP) as a portrait of a news media outlet's attention to various nations. GAP software automatically crawls a news media outlet's website and calculates country-by-country story counts over a period of time. This paper reports these story counts and correlates them to a wide range of country data sets provided by the World Bank.

GAP research demonstrates that the most accurate predictor of a media outlet's attention is the size of a nation's gross domestic product. This correlation is significantly greater than the correlation between media attention and the size of a nation's population, and appears to be the strongest correlation between media attention and 21 factors examined. Generally speaking, violent conflict seems to have less effect on media attention than the size of a nation's economy does.

While most media sources studied demonstrate similar patterns, one media outlet – the BBC News – shows radically different patterns. The BBC's media attention is more closely correlated to the size of a nation's population than to the size of its economy.

Introduction

On April 7, 2003, the New York Times carried a story on a massacre in the Democratic Republic of Congo in which up to 966 non-combatant civilians were slaughtered by warring factions in the east of the country. The story, authored by the Associated Press, ran on page A6 and took up less than a full column of newsprint. Amid the flurry of Iraq war coverage (to which the New York Times devoted no fewer than 5 cover stories and an entire special section) many English-language newspapers made no mention of the Congo massacres; nearly all that did ran a short excerpt of the AP story.

It seems unlikely that the New York Times would give similarly cursory treatment to a massacre of 966 civilians in France, Britain, South Korea or Israel. This disparity leaves us with a question: Is Iraq more important than the Congo?

While these may be uncomfortable questions, they are ones asked and answered – implicitly – journalists and editors every day. When journalists decide to report on one story and not another, they decide what their readers will pay attention to. When editors assign reporters to certain countries, they decide what nations will receive major coverage and which will be covered cursorily. When news media outlets "embed" more than 600 reporters in Iraq[1], they are telling their readers that the US invasion is more important that the long-running multinational war centered in the Democratic Republic of Congo.

For an "apples to apples" comparison, it is useful to consider whether Japan or Nigeria is more important. Their populations are roughly equal – 130 million in Nigeria, 127 million in Japan. Neither is short on possible news stories. Nigeria, in particular, seems to have all the factors we commonly associate with headline news: crime, violence, ethnic strife and religious conflict.

If we define "media attention" as "the number of stories on a given subject", the statistics give us a clear answer: Japan is roughly seven times more important than Nigeria. Searching the archives of seven media sites and two media aggregators, we find between 2 times (BBC) and 16 times (CNN) as many stories that reference the search string "Japan" as those that reference the search string "Nigeria", averaging 7.28 times as many Japanese stories across the sources sampled.

	AP	AltaVista	BBC	CNN	Google	NYPost	NYTimes	Reuters	WPost	
Japan	362	191728	2589	9863	16800	3119	712	411	63	
Nigeria	53	25361	1385	623	5830	328	119	42	12	
Japan multiplier	6.83	7.56	1.87	15.83	2.88	9.51	5.98	9.79	5.25	**7.28**

[1] "'Embedded' reporters are mixed blessing for the military", Josh Getlin and Tracy Wilkinson, http://seattletimes.nwsource.com/html/television/134667871_embed03.html, accessed July 31, 2003.

A comparison of these two countries challenges some conventional answers we might give to the question, "What does the media pay attention to?" We might expect English-language media to author more stories on English-speaking nations. However, Nigeria's English-speaking and Japan is not. Are our media outlets more likely to report on stories close to home? Nigeria is closer to the headquarters of each of the individual media sources than is Japan. Does our media report on "people like us" – i.e., people with whom we share a religious or ethnic background? Japan's Buddhist homogeneity has less in common with the USA's Christian-heavy religious pluralism than does Nigeria's 50/40 Muslim/Christian split[2]. And the United States has many more African-Americans than Asian-Americans.

So why do media outlets pay so much more attention to Japan than to Nigeria? One possible answer is "economics". Nigeria's 2001 GDP was $41 billion, making it the 54th largest economy in the world, ranking between Bangladesh and Libya. Japan's 2001 GDP was $4 trillion, second only to the United States, the size of France's, Britain's and China's economies combined. As discussed in this paper, the distribution of attention in a single media source is more directly proportional to national GDP than to any other single factor – this fact goes a long way towards explaining the Nigeria/Japan attention disparity.

The GAP model is designed to examine a media outlet, or set of media outlets, and answer two questions: On which countries does this media outlet focus its attention? How does this attention distribution correlate to a wide range of factors? The GAP software polls the search engines of media sites with very simple requests – in most cases, the name of each of 183 nations – and compares the resulting distribution with other widely available data, primarily development data from the World Bank.

Why should one care about media attention? Several reasons seem apparent:

- **Trade** – In a globally interconnected economy, there are at least indirect economic consequences to the distribution of attention. As trade becomes global, it becomes crucial for nations to be globally visible as possible trading partners. India's IT revolution has been a triumph of both education and marketing – not only have India's universities developed tremendous capacity for training top IT professionals, India has also "branded" Bangalore and Hyderabad as world-class IT centers. As a result, multinational corporations have felt comfortable outsourcing major IT projects to Indian firms, spurring a high-value industry. Some middle-income nations have been engaging in branding that is almost corporate, producing inserts for magazines like Newsweek International to promote their nations as product. In part, GAP attempts to look at how successful different nations have been at "getting their brand out".

- **Aid** – Individuals, NGOs and governments contribute small, finite sums in the form of humanitarian assistance to developing and conflict-ridden nations. This

[2] CIA World Factbook, http://www.cia.gov/cia/publications/factbook, accessed July 31, 2003.

money has a tendency to flow towards the conflict most visible at any particular moment – one might term this the "Live Aid" effect. Nations with less well-publicized needs tend to go wanting. After US intervention in Afghanistan, substantial commitments were made by organizations and governments to rebuilding that nation. At the time, many international aid groups expressed concern that other nations were also in need of assistance and that aid to Afghanistan – the high-profile conflict – might detract from aid to other nations. Now that Afghanistan is no longer as prominent in the global media, it appears that many pledged funds will not be forthcoming. Afghanistan may find itself short on reconstruction funding as those new funds head to Iraq, today's high-attention country.

- **Intervention** – The more attention that is devoted to a particular conflict, the more likely it is to attract foreign aid and, in rare cases, military intervention. Individual nations and multilateral coalitions have a tendency to intervene in high-visibility conflicts and to ignore conflicts in less visible nations. (Possible US intervention in currently-high-visibility Liberia on humanitarian grounds would contrast with the lack of planned US intervention in low-visibility, but severe, conflict and human rights violations occurring in Sudan.) Global media attention makes it more likely that the United Nations and other multilaterals will intervene to prevent or halt genocides. Global media attention likely prevented many deaths in the mid-1990s Balkan conflicts (after the low-attention deaths of many others), while a near-total lack of media attention marked the massacres in Rwanda in 1994, which occurred largely without outside intervention.

On a more individual level, media attention is important because informed decisions and opinions require factual input. Whether one considers the recent war in Iraq to be a victory for democracy or a tragic triumph of unilateralism, it is fairly easy for citizens to have and support their opinions in the wake of tens of thousands of stories written about the conflict.

It is much harder to have an educated opinion about whether the Hema or the Lendu are in the right in their conflict in Bunia, DR Congo, for the simple reason that very little, comparatively, has been written about the conflict. The New York Times's A6 story on the Congo massacres is a case in point.

To evaluate the success or failure of the key media news outlets in informing their readers about the state of the world, more is required than an intuitive sense that certain stories are going unreported and certain nations are being ignored – what is needed is quantitative evidence. The GAP methodology set forth in this paper is an attempt to supply this evidence.

Methodology

The core of the Global Attention Profile project is a set of Perl scripts – scrapers – that query the search engines of nine new media outlet websites and perform 183 automated searches. The scrapers collect a single piece of data from each search – the total

number of stories available within a given time period for a given search term – and present this data in an HTML table and on a world map. Using previously calculated equations, the scrapers estimate how many stories "should" result from a given search term, estimating based on a nation's population and GDP. The program calculates the deviation between the two predictions and observed results, and reports this deviation on the HTML table and on maps.

Output of each script is an HTML table and three maps – hitcount, deviation from GDP prediction, and deviation from population prediction. Scripts can be run at an arbitrary interval, called on a Unix system via crontab. Scripts have run daily for the past three months – in the future, they will likely run on a weekly basis, as data changes little on daily intervals.

Scrapers

"Scraper" is a generic term for a program that requests a webpage, selects certain data from it and returns that data in a different format. Before the advent of syndication formats like RSS, programmers routinely used scrapers to retrieve news headlines from multiple sources and aggregate them into personal news sites for instance. In this case, scrapers make it possible to rapidly query search engines and select one piece of data from the results – the total number of results the search engine links to. The scraper creates a custom URL – the general URL for querying a specific search engine, plus a query term corresponding to the nation we're searching for – and, because the website in question believes it is responding to a request from a web browser, receives an HTML file in response. The scraper then uses regular expression matching to retrieve the string that contains the total response count.

GAP scrapers query seven websites: news.google.com, www.AltaVista.com/news, query.nytimes.com, pqasb.pqarchiver.com/nypost, www.bbc.co.uk, search.cnn.com, and www.washingtonpost.com, which is queried for AP and Reuters results, as well as for Washington Post results. Because there is no industry-standard search engine, or universally accepted search protocol, it is necessary to approach each engine slightly differently, using four data files, which differ only in how they pass boolean queries to the engine (NOT represented as "-", "NOT", "AND NOT" or no support for NOT). A unique configuration file was used for each engine, which tells the scraper which data file to use, as well as the base search URL and the regular expression that matches total results.

The first edition of the GAP scraper was written by Chris Warren; the author is responsible for subsequent versions.

Search Terms

The goal of GAP is to compare the representation of different nations by a news media outlet; the first challenge was finding search terms that generate stories on a given nation. For the purposes of GAP, nations and territories with a population over 100,000 and current population and GDP statistics were most interesting.

Does a search for ***Argentina*** return all the Argentine-related stories within a collection? Obviously not. Stories that reference ***Buenos Aires***, but not ***Argentina***, will be skipped, as will stories that refer to ***Argentines***. Sometimes a search may be overbroad – a search for ***Tonga*** will match stories listing football teams playing in the World Cup. Is that story about Tonga or not?

Worse still, there are the inconveniently-named nations. Searching for ***Chad*** will find webpages on country singers and baseball players before identifying the African nation, and ***Georgia*** will net far more articles on Atlanta than on Tbilisi.

GAP acknowledges all of these difficulties and then ignores most of them. Automatically determining the topic of a piece of text is one of the most difficult problems in computer science. The best automated systems rely on human judgment to construct a corpus of hand-sorted documents for the system to "learn" from. In other words, there's no way to rapidly determine a document's subject at a high degree of accuracy without both human judgment and expensive, complex software.

That is not to suggest that the simple names of nations are the ideal terms for GAP – a future version may use a string like ***"Great Britain" OR England OR "United Kingdom" OR Scotland OR Wales OR London OR "Downing Street" OR Welsh OR Scotch OR English OR British*** to match for the United Kingdom. Given the existence of cities named London in both the US and Canada, and of Scotch whisky, the solution may create more problems than it solves.

GAP tries to address the most egregious problems through the judicious use of quotes and boolean search terms to constrain overbroad searches. In the cases of Georgia, Chad and the Republic of Congo, it searches for the name of the national capital, rather than the name of the nation, and compensates by multiplying the number of returned searches by five. (Searches of the twenty nations and capitals closest to each nation in terms of total GDP set 5x as the appropriate multiplier.) GAP ignores the United States altogether, given massive undercounting of stories set in major US cities.

GAP strives to be comparatively accurate, not absolutely accurate. If GAP reports that news.google.com turns up 15,000 results for ***Japan***, one should not conclude that there are 15,000 stories on Japan – there are likely more, and possibly fewer. However, GAP tries hard to be consistently inaccurate, so that when comparing 15,000 results for ***Japan*** and 3,000 for ***Nigeria***, it's reasonable to say that there are five times as many stories on Japan than on Nigeria.

Search Engines

An ideal search engine, for GAP purposes, would support boolean queries, interpret quoted strings as literal strings, give exact and verifiable numbers of total results and allow any date range to be queried. Unsurprisingly, the ideal engine does not yet seem

to exist[3] – in every case one must compromise, somewhat, making "apples to apples" comparisons of results inexact. The table below summarizes the characteristics of the seven sites queried:

Site	Exact	Verifiable	Boolean	Quote	Date
AltaVista	Yes	No	Yes – minus	Yes	7,30, range
BBC	No	Yes	No	Yes	Full – 1997
CNN	Yes	No	Yes – NOT	No	Full – 1996
Google	Yes	No	Yes – minus	Yes	30 days
NY Post	Yes	Yes	Yes – AND NOT	Yes	2 years, Full – 1998
NY Times	No	Yes	Yes – NOT	Yes	7,30,90, 365, Full – 1996
WPost, AP, Reuters	Yes	Yes	Yes – NOT	Yes	1-14 days

Exact – To perform meaningful comparisons of results counts requires an exact number, rather than a range. Many engines refuse to give an exact count, offering a string like "more than 1000 matches" for large queries. BBC does not offer a count of stories when it matches fewer than 10 results – the next version of the GAP scraper will accommodate this special case, but the current one does not.

While some problems can be worked around (the "more than" problem can be defeated by specifying a short date range, for instance), others are insurmountable: Fox News Channel (foxnews.com) returns three results for each query, and an additional ten on follow-on pages, but never tells you how many results or pages are available. Evidently "we report, you decide" doesn't apply to quantitative analysis of their news coverage.

The New York Times appears to provide an exact count, but the count is not believable for date ranges above 90 days. Year-long and full archive searches never return more than 1,000 results, implying a manual trimming of archives. 90-day counts appear believable – they are roughly three times the size of 30-day counts.

Verifiable – When a search engine reports 980 results, users expect to be able to view any of those 980 pages. In three cases, this isn't possible. CNN will only provide access to the top 500 stories it matches for any query. While other stories may be there, it is possible only to verify the first 500.

AltaVista and Google, the only two news aggregators in the set, have major verifiability problems. Google will routinely report 45,000 results for a search. When one pages through search results, as few as 1% of the stories will be user-viewable – in other words, there will be 45 pages of results, rather than the 4,500 one would anticipate.

3 Commercial news aggregators like Lexis/Nexis come close to being this ideal search engine. Since they log all stories into their own database, one can search across multiple media outlets with a common set of keywords and time periods. Once GAP is modified to search a news aggregator, much of the complexity detailed in this section will be extraneous.

AltaVista shares this limitation, though camouflages it – request a high-numbered page of search results on AltaVista and you'll get the first page of results!

Google has another peculiarity – changing story numbers. While the number of total results returned by news.google.com is constant over short periods of time, the number of stories a user can view varies from moment to moment. The variance is small, under 5%, but it implies that Google is querying one server for a total story count, and others for the story summaries.

These phenomena may be explainable as byproducts of search engine optimization. While many users rely on result count as a measure of a search's exactness, very few request the 1000th story returned for a particular search – as a result, the engine is optimized to provide the first piece of data but makes it difficult, if not impossible, to retrieve the second piece of data. (One might also conclude that AltaVista and Google have done this to prevent scrapers and bots from spidering their site by performing broad searches and collecting all the URLs referenced in results.) In other words, while unverifiable results don't necessarily imply incorrect results, they are a cause for concern in collecting valid data.

Boolean – A search for ***Ireland*** is likely to return stories on the Republic of Ireland (the target) and on Northern Ireland, which is part of the UK and hence not part of the target. It's useful to be able to ask a search engine for ***Ireland NOT Northern***. Indeed, it becomes very important when searching for information on Guinea, which tends to turn up matches on Guinea-Bissau, Equatorial Guinea and Papua New Guinea, not to mention the Gulf of Guinea, guinea hens and guinea pigs.

BBC is the only engine queried that does not support boolean searching. As a result, certain results (Guinea, Congo, Niger, Ireland and others) are bound to be overbroad. At the moment, GAP does not compensate for these overbroad matches.

Quotes – A search for the string ***"west bank"*** should return stories about the middle east, while a search for the string ***west bank*** should give us all those stories, plus stories about the ***bank*** opening on the ***west*** side of town. CNN does not recognize quotes, and hence searches like ***"Equatorial Guinea"*** are overbroad. Again, GAP does not compensate for these overbroad matches.

Date – To compare "apples to apples" between data sources, one should look at the same time period for both sources. Unfortunately, search engines vary widely in what date ranges can be searched. Three engines – New York Post, BBC, CNN – provide only multi-year searches. The Washington Post data sources, which include AP and Reuters, are only searchable for the past 14 days. And the New York Times, while quite flexible in time periods permitted, doesn't allow a 14 day search (for easy comparison with the Washington Post), and doesn't provide believable results for periods over 90 days.

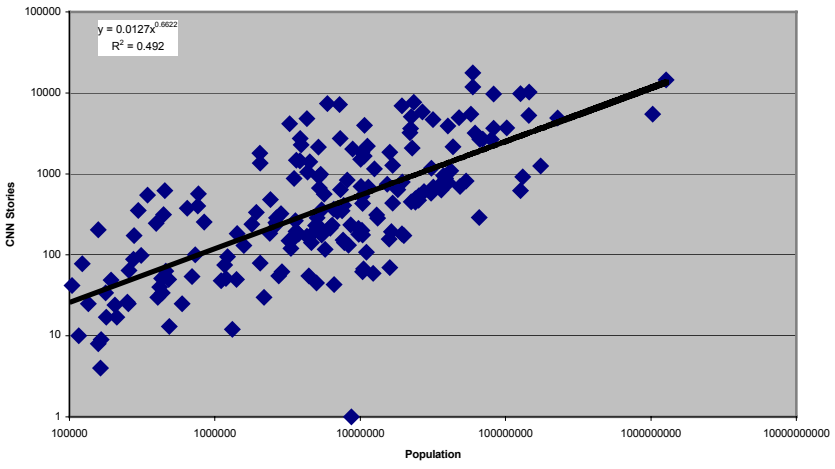
Differing date ranges means it's impossible to compare CNN and Google without considering the time scope – any comparison needs to recognize that Google measures a short slice of time compared to CNN's multiyear swath. Short-term phenomena – the war in Iraq, for instance – are likely to affect a month-long sample more dramatically than a multi-year sample.

Correlations and estimators

The results of the GAP scraper, by themselves, are relatively unhelpful. What does it mean that CNN has 629 stories on Sudan? Is 629 a lot or a little?

GAP attempts to contextualize by building two models for the distribution of data and comparing actual results to the models. One model is built on population data, the other on GDP. These two statistics are used because they are ones for which the most thorough data sets exist – the World Bank provides 2001 data[4] on population for but one country (Mayotte), and GDP data on 90% of our countries. For the remaining nations, the CIA World Factbook's 2000-2002 estimates[5] were used.

To create an estimator model, a correlation between population and results count was assumed. The logarithm is taken of both population and results counts, which results in distributions that appear to match a normal distribution for each data set. The log of story count was graphed against the log of population on a scatterplot, and a line was fit to the results. This linear fit to logarithmic data is equivalent to fitting a curve of the form $y=mx^n$ where y=stories, x=population and m and n are constants specific to that particular data distribution.[6]



The chart, left, shows the relationship between search results on CNN.com on June 11th, 2003 and population. Both axes are scaled logarithmically, and as a result, the curve $y=0.0127x^{0.6622}$ appears linear. The equation was used as estimator for future CNN results. When

[4] From the World Development Indicators Database, via the Data Query online tool: http://www.worldbank.org/data/dataquery.html, accessed June 15, 2003.

[5] CIA World Factbook, http://www.cia.gov/cia/publications/factbook, accessed July 31,2003. Unfortunately, the CIA World Factbook's estimates of GDP are in purchasing power parity dollars, not real US dollars. As there's no easy way to convert PPP to real dollars, the use of Factbook introduces some error into GDP comparisons.

6 Indeed, when Microsoft Excel fits a power series to a set of data, it appears to take the logarithm of both sets and fits a linear equation to the transformed data. As a result, Excel is unable to perform power series fits to data that includes zeros, as it's impossible to take the logarithm of zero.

m and n from this equation and from the corresponding GDP equation are plugged into a CNN specific configuration file, the scraper is able to estimate how many results it expects based on the population and GDP models, and calculate deviation from those expectations. For ease of visualization, it color-codes the deviations to make maps and charts easier to read, using deeper shades of red to signify greater positive deviations (more results than the model predicted) and deeper shades of blue to signify negative deviation (fewer results than expected.)

It is tempting to read blue spots on the map as "underrepresented" and red ones as "overrepresented", but these generalizations have to be made with a strong caveat. Blue and red signify under and overrepresentation *from a specific model*. As will be demonstrated in the **Results** section, models are more or less well-correlated to a specific data set, and one would expect large amounts of deviation from a loosely-correlated model.

Results were then correlated from the nine media sites with 21 World Bank data sets[7]. With the exception of the aforementioned GDP and Population statistics, these sets represent 1999 data. Few are as complete as the GDP or Population statistics, so correlations consider 120-160 pairs of values rather than the 183 considered by GDP and Population correlations[8]. It was not assumed that missing data indicates a zero value – generally this is untrue, and would badly skew correlations. In the case of Foreign Direct Investment and Development Assistance, negative values – i.e., countries making investments or countries giving aid – were disposed of, because this information is extremely incomplete.

Microsoft Excel was used to fit curves to the data. Because Excel transforms data via a logarithm to effect a power-series curve fit, it is not capable of working with zero values. To work around this problem, zero values returned by our scrapers were replaced with 0.1 – one tenth of a story. As this study makes use of logarithmic scales, this change turns the difference between zero and one into a difference of one order of magnitude, rather than an infinite difference – probably a better representation of what is actually going on, especially on news sites that might fail to report on Vanuatu this week, but provide a story on it next week.

Visualization

Some of the most interesting patterns that GAP reveals are geographic. For instance, it is easier to see that the BBC focuses a lot of reporting attention on former British colonies in east and southern Africa once results are plotted on a map. For this reason, the GAP scraper outputs three maps, as well as a table of values. The maps represent results count in percentage terms, deviation from population estimate and deviation from GDP estimate.

[7] World Development Indicators Database.

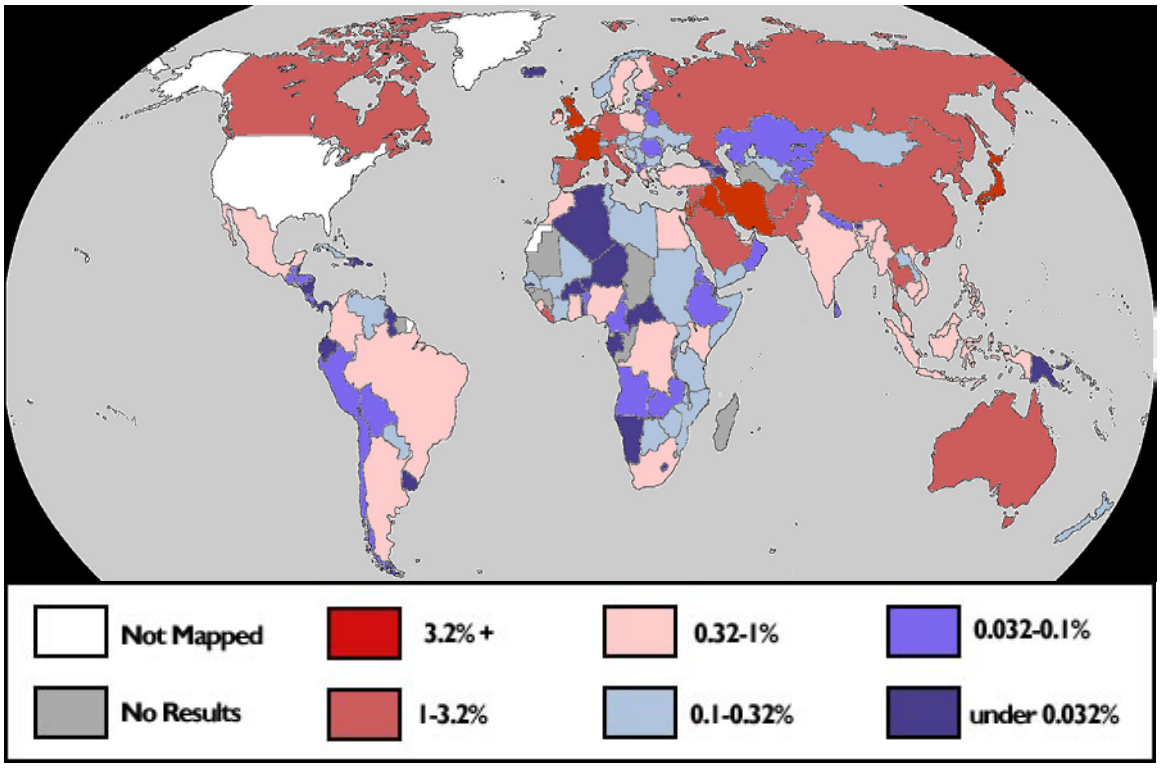
[8] As a result, comparing correlation coefficients is somewhat imprecise, as a data set with 120 data points will show slightly different correlations than one with 180 data points.

To automatically create these maps, the GAP scraper calls on mapper, a package built by programmer Nate Kurz to enable automated mapping using ImageMagick, a leading open source imaging tool. Mapper reads a data file containing x,y coordinates for each nation represented on a specific map. Some discontinuous nations (Indonesia, for instance) require multiple coordinate points. Mapper defines a region as being represented by one or more coordinate pairs, and then accepts commands to fill a region with a certain color, coded in hexadecimal RGB pairs and outputs PNGs, GIFs or JPEGs.[9]

Results

GAP's first aim is to provide a picture of a media source's attention profile on a given day. Because the scrapers are constrained by the date range of the given search engine, a given picture might represent a time period from the past 14 days to the past several years.

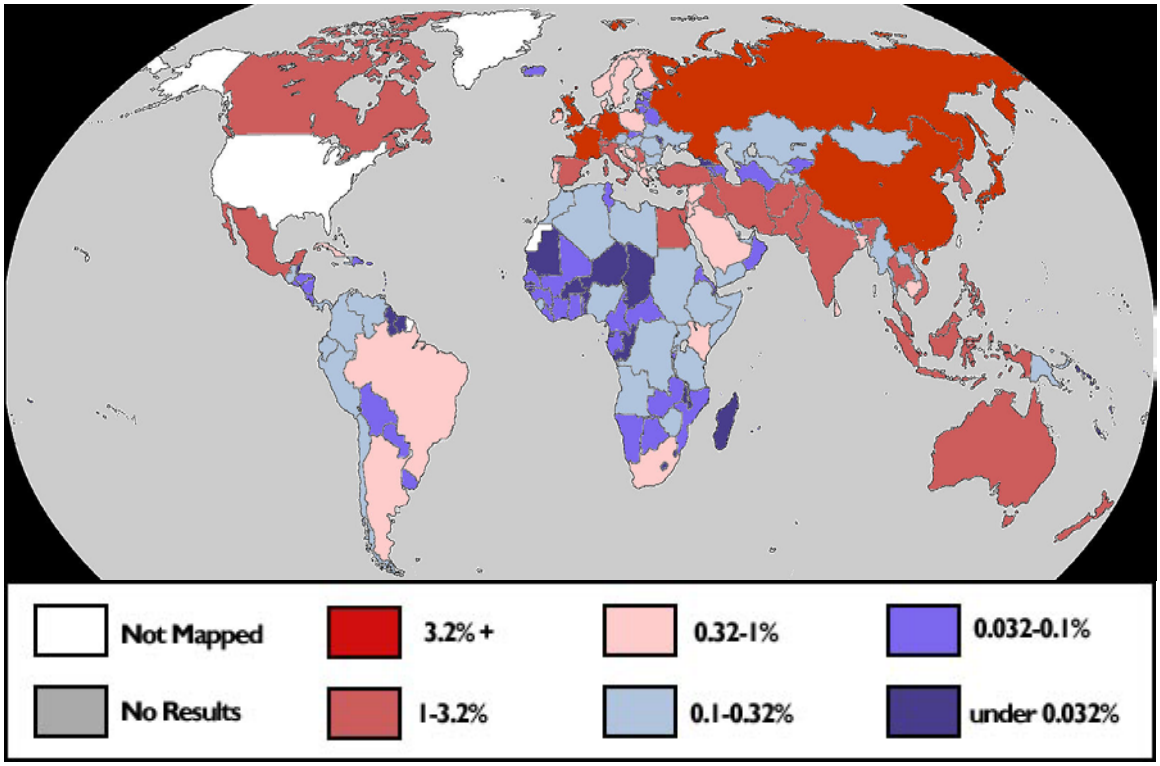
The following is a picture of Reuters' attention profile for June 11th – 25th. The coloring of the map represents what percentage of stories detected by the GAP scraper reference a particular nation. A search for ***Iraq*** turns up 1,352 stories, of 13,360 total retrieved in this time period, or 10.11% – as a result, Iraq is colored bright red. ***Algeria***, by contrast, retrieves 2 stories, or 0.015%, and is colored deep blue. In the two week period, there are no stories about Mauritania, Turkmenistan, Madagascar and a few others, so they are colored grey.



9 The author and Mr. Kurz plan to release mapper to the Open Source community at some point in the future. If you need the code in the meantime, please contact ethan@geekcorps.org.

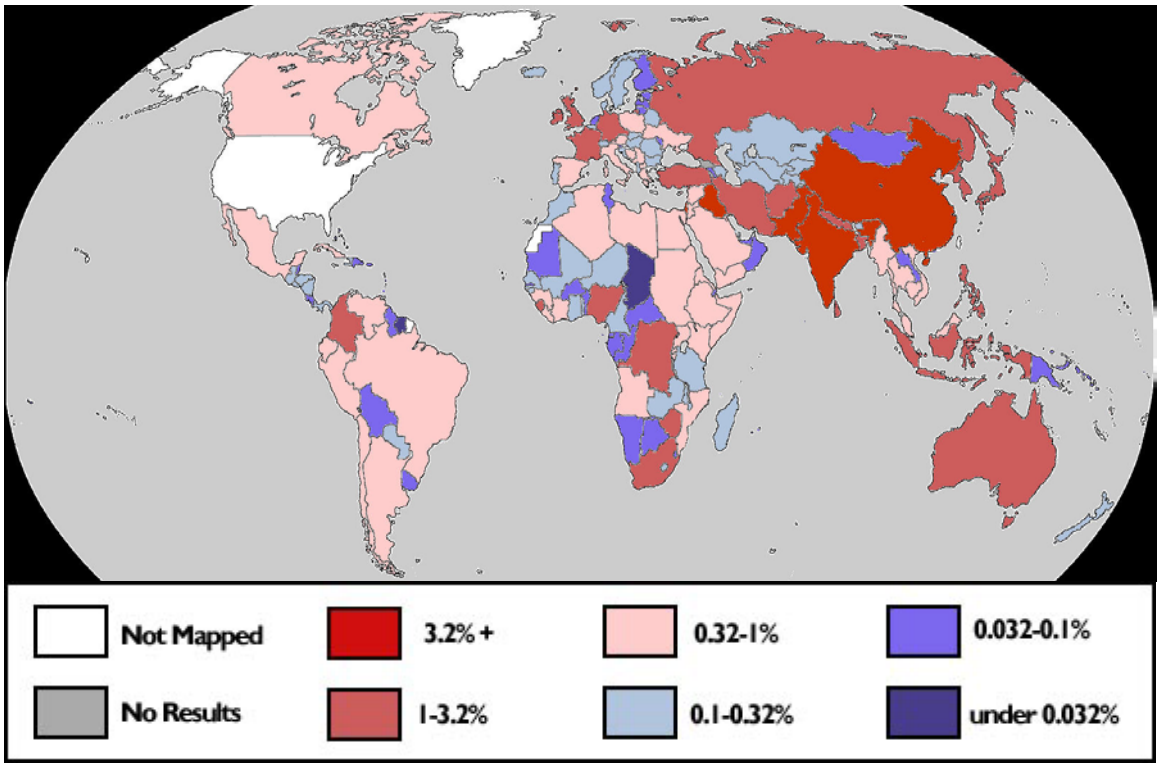
A map for this brief a time period does a good job of revealing breaking stories. Liberia and neighboring Sierra Leone stand out in red, against neighbors in blue and grey, due to rebel activity in Liberia, some from bases in Sierra Leone. The area around Iraq is still bright red, in the aftermath of the US/UK invasion. These maps change quite quickly – a map taken two weeks later will use data that has no overlap with data plotted in this map, and it's likely there will be some major coloration change. (Readers can check – a map approximately two weeks later is available online at http://h2odev.law.harvard.edu/ezuckerman/maps/reuthits20030711.jpg)

Maps of longer time periods are useful to get a clearer sense for overall media trends. The following map of CNN represents stories from 1996 to the present. As a result, it does a poor job of showing current stories, but a better job of showing overall patterns of coverage.



Generally speaking, coverage is concentrated in Western Europe, the Middle East and Southeast Asia, with good coverage in the large economies of China, Japan, Mexico and Canada. There is very little coverage in most of Africa (Kenya, with the 1998 US embassy bombings, and South Africa, the largest economy in the region are exceptions), in Central Asia, Eastern Europe and in most of Central and South America (the large economies of Mexico, Brazil and Argentina are exceptions.

A map of BBC coverage for a similar time period (1997 – present) contrasts sharply:



The pastel tones imply a more even media distribution than on the CNN map. (If each country got the same number of stories, each would have 0.54% of stories and would be colored light pink.) Africa is a major contrast – while French-speaking parts of West Africa are blue, the English speaking parts of West Africa, as well as most of East and Southern Africa, are well covered. Central Asia and Central America are still sparsely represented, and there is a surprising blue patch over Scandinavia, better represented on the CNN map. (It is possible that some of the low counts in Western Europe are a result of BBC's tendency to refer to European cities without mentioning the country they're located in, something American media sources do less frequently.)

Such different maps suggest that BBC and CNN have different criteria for story selection, place reporters differently, and generally have a different way of paying attention. Correlating story counts to population and GDP bears this suspicion out. CNN shows correlation to population (R^2=0.49), but much stronger correlation to GDP (R^2=0.69). BBC is just the opposite – it is loosely correlated to GDP (R^2=0.38) but tightly correlated to population (R^2=0.67)[10].

The maps thus far speak volumes about how stories are distributed, but not how they *should* be distributed. In every map generated thus far, Iraq has at least 1% of total stories, more than 3.2% in two of the three maps. Is Iraq receiving more attention than

[10] It is unlikely that BBC consciously chose for its coverage to tightly track population distribution, just as it is unlikely CNN chose to closely track capital distribution. It's more likely that BBC has an unstated policy of closely following former British colonies, which keeps it focused on Africa and South Asia.

one would generally expect, due to the recent war, or is Iraq sufficiently important to warrant this attention?

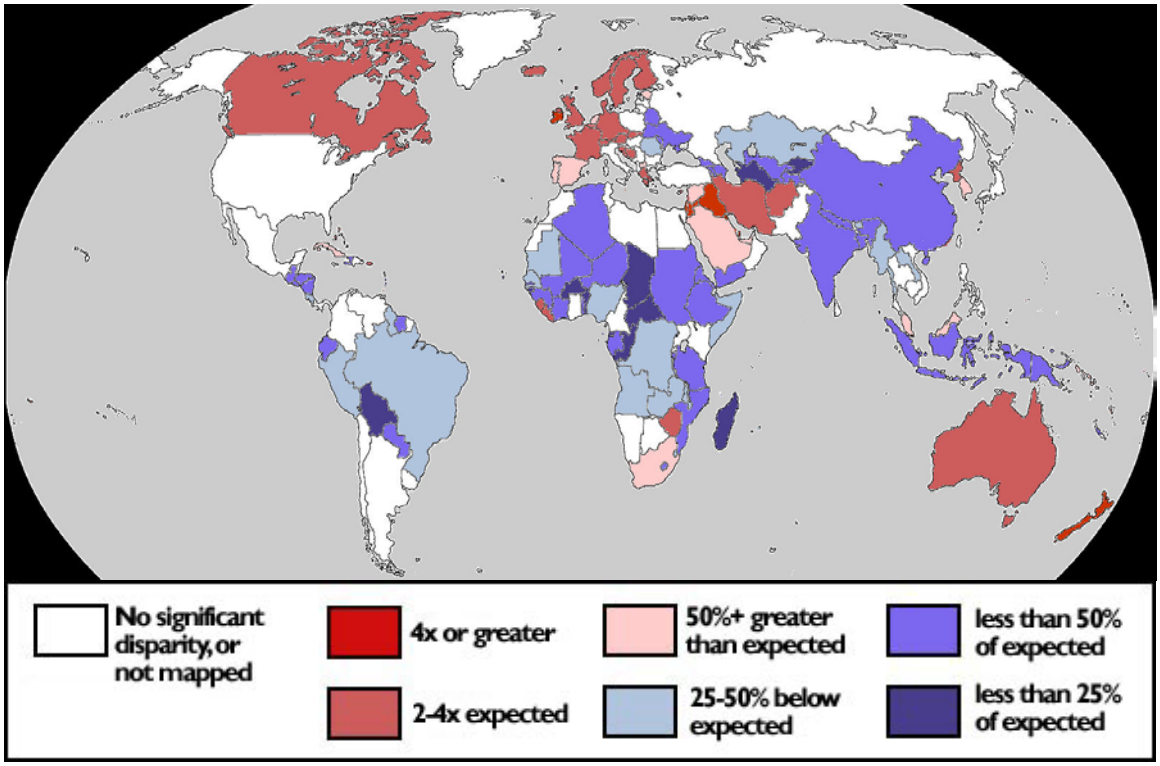
To answer this question, GAP estimates likely story distributions, extrapolating from actual story distributions.

An extremely naïve estimator model would make the assumption that every nation should receive the same amount of attention. Thus, one would assume each nation should have 0.54% of retrieved stories, and would mark nations that received more as "high attention" and those receiving fewer as "low attention". This model does not stand up to close examination, though – is it really reasonable to expect Tonga to receive as much attention, with a population of 100,000, as China would with a population of 1.3 billion?

Acknowledging this problem, one might advance the "Andy Warhol model" – an assumption that everyone will receive 15 minutes of fame – and assume that story distribution would be directly proportional to population distribution. For this to hold true, every story on Tonga would be counterbalanced by 12,700 stories on China. Obviously, this is not the case – even a small nation like Tonga appears periodically in mainstream media, if only to acknowledge its participation in UN votes or international rugby matches.

To create a less naïve population-based estimator model, actual results from the scrapers are examined, to look for correlation between population and story count. On news.google.com, a loose correlation exists (R^2=0.45) between population and story count. Using the equation from the best fit curve, one can speculate what a story distribution would be **if** story count and population were perfectly correlated. The next step is to compare actual distribution to this estimation and map the differences. (This process is described in more detail in the **Correlation** subsection of the preceding **Methodology** section.

Here is the resulting map for news.google.com on June 27, 2003

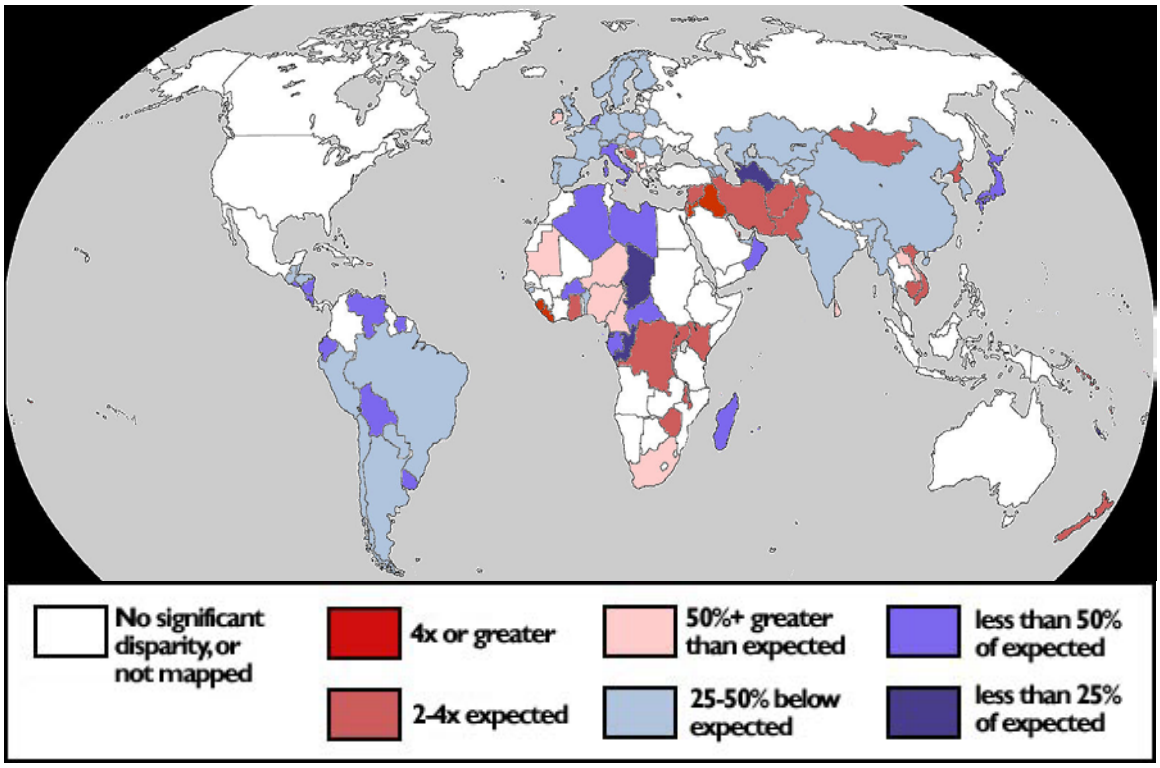


Western Europe, Australia, New Zealand and Canada are in shades of red – each has a comparatively small population but receives a good deal of media attention, generally 2 to 4 times what one would expect based on their population. The Middle East, the Korean Peninsula and a few African nations appear in red, probably due to breaking news (the ongoing violence in Liberia, Mugabe's struggle for power in Zimbabwe, North Korea's nuclear threats).

Most of Central and South America are blue, as is most of the African continent, Eastern Europe and Central Asia. Some countries receive fewer than 1/4 of the coverage one would expect based on their population. China, Indonesia and India, three of the four most populous nations, show up medium blue – with such large populations, they would need a large number of stories to meet their expected distributions.

If one expects the 4,500 news sources tracked by Google News[11] to represent the world's population evenly, this map suggests imminent disappointment, especially if one is searching for news on poor countries. Given this map's resemblance to a map of GDP per capita (rich nations in red, poor ones in blue), a logical next step is to build a second estimate based on national GDP. Using the same technique, the following map is generated, representing deviation of actual Google News results from a GDP-based estimation on June 27, 2003:

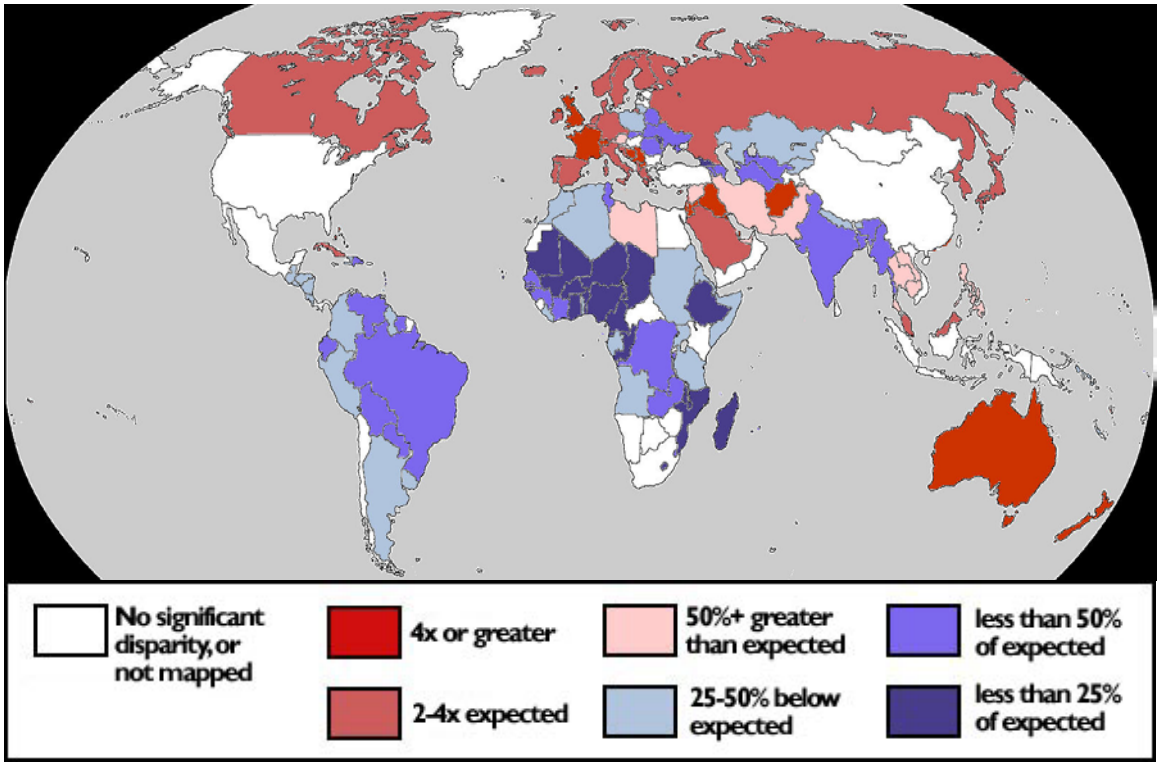
[11] "Google News (Beta)", http://news.google.com/intl/en_us/about_google_news.html, accessed July 31,2003.



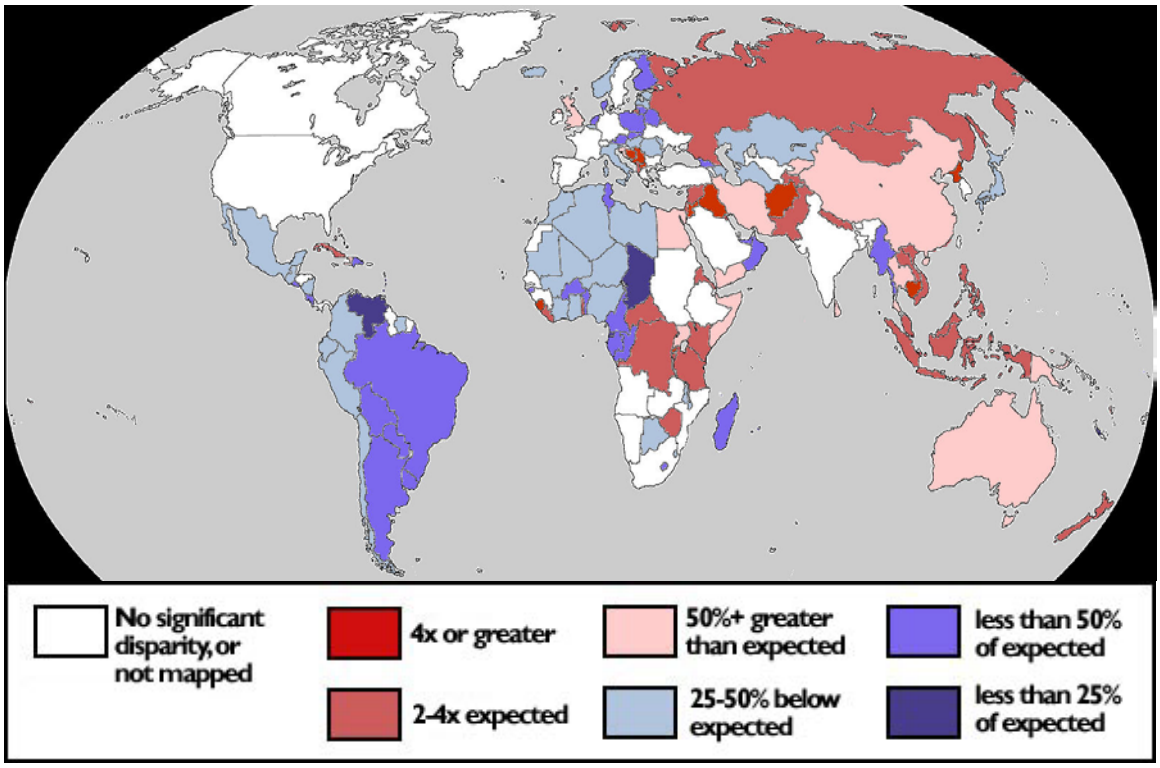
The predominance of white and pastel colors reflects the fact that GDP is far more closely correlated to Google News results than population (R^2=0.62 versus R^2=0.45) Several large economies – Western Europe, Japan, South Korea – suddenly appear as underrepresented, while a number of African nations register as over-represented. Central Asia and South America remain blue, less represented than would be expected in either GDP or population terms.

CNN's GDP and population maps give one a sense for how these variations play out in the long term, as CNN results represent over half a decade of data.

CNN variation from population estimates, June 27, 2003:



Over the long term, Africa, South and Central America, Eastern Europe and Central Asia receive less attention than predicted, while Western Europe, the Middle East, Russia and Oceania receive more than predicted. The picture in terms of GDP is quite different:



While West Africa still goes largely unwatched, Southern and Central Africa receive attention disproportionate to their economies. Parts of Central Asia now see attention proportional to their GDP, if not to their population. Southeast Asia also receives more attention than would be predicted, while parts of Western Europe receive less than anticipated.

Largely unchanged between the two maps is the Middle East, better represented than one should expect based on either population or GDP. South and Central America go underrepresented by both estimations. Especially interesting are Brazil and Argentina, large countries (5th and 31st in 2001 population) with large economies (11th and 17th respectively in 2001 GDP).

Since neither population nor GDP gives a fully accurate estimate of story distribution, one may ask whether any other factor provides a better picture of how stories are distributed. To answer this question, the results from all nine scrapers were correlated with 21 data sets provided by the World Bank. A chart below summarizes the correlations:

Values shown are the value of the squared correlation (R^2) between the data set and the power series regression equation. For all correlations, $p<0.0001$.

	AP	AltaVista	BBC	CNN	Google	NYPost	NYTimes	Reuters	WPost	Average
GDP	0.53	0.66	0.38	0.69	0.62	0.64	0.66	0.52	0.53	0.58
Goods and service imports	0.53	0.67	0.31	0.69	0.64	0.62	0.66	0.52	0.53	0.58
Total PCs	0.53	0.66	0.35	0.69	0.62	0.64	0.64	0.49	0.53	0.57
Urban Population	0.50	0.55	0.62	0.58	0.53	0.49	0.59	0.50	0.53	0.55
Goods and service exports	0.50	0.64	0.29	0.66	0.61	0.58	0.62	0.46	0.49	0.54
Military personnel	0.41	0.45	0.58	0.56	0.48	0.41	0.50	0.45	0.44	0.47
Internet Users	0.44	0.58	0.24	0.59	0.55	0.53	0.52	0.40	0.44	0.48
Population	0.42	0.45	0.67	0.49	0.45	0.37	0.50	0.44	0.44	0.47
Mobile Phones	0.42	0.55	0.26	0.53	0.52	0.50	0.53	0.45	0.42	0.47
Literate Population	0.40	0.44	0.64	0.49	0.46	0.38	0.48	0.38	0.46	0.46
Aircraft Departures	0.37	0.52	0.23	0.53	0.50	0.42	0.42	0.35	0.41	0.42
Kilometers of road	0.37	0.41	0.45	0.44	0.41	0.36	0.44	0.36	0.34	0.40
Foreign direct investment	0.35	0.46	0.14	0.45	0.41	0.46	0.42	0.34	0.37	0.38
Tourism, arrivals	0.34	0.44	0.13	0.42	0.42	0.42	0.44	0.34	0.35	0.37
Currency transfer from abroad	0.32	0.34	0.24	0.36	0.32	0.40	0.42	0.24	0.32	0.33
Tourism, receipts	0.27	0.44	0.08	0.41	0.37	0.41	0.38	0.30	0.35	0.33
Arable Land	0.28	0.28	0.50	0.29	0.29	0.24	0.31	0.29	0.30	0.31
Workers remittances	0.20	0.30	0.27	0.35	0.31	0.48	0.35	0.20	0.23	0.30
Surface area	0.21	0.17	0.40	0.21	0.18	0.14	0.23	0.20	0.19	0.22
Freshwater resources	0.07	0.08	0.19	0.14	0.08	0.07	0.09	0.06	0.12	0.10
Development assistance	0.09	0.07	0.14	0.08	0.08	0.07	0.10	0.10	0.10	0.09

Five of the factors – total GDP, imports and exports of goods and services, total personal computers nationwide and urban population – correlate well with scraper results: their squared correlation(R^2) is 0.5 or better, which means that more than half the data distribution is explainable by the correlating equation. Seven factors – military personnel, Internet users, population, literate population, aircraft departures and kilometers of road – are loosely correlated to scraper results: their R^2 correlation is above 0.4. The remaining nine factors are probably not correlated to article distribution as reported by scrapers.

GDP and Goods and Service Imports tie for highest correlation, with R^2=0.58 on average. All data sets except BBC are most closely correlated to either GDP or goods and service imports – BBC, alone, is most closely correlated to population.

Dividing the 21 World Bank data sets into five categories – Economic Indicators, Population Indicators, Technology Indicators, Globalization Indicators and Physical Indicators – helps provide a sense for what types of data correlate most closely to story distribution:

	AP	AltaVista	BBC	CNN	Google	NYPost	NYTimes	Reuters	WPost	Average
Economic indicators										
GDP	0.53	0.66	0.38	0.69	0.62	0.64	0.66	0.52	0.53	0.58
Goods and service imports	0.53	0.67	0.31	0.69	0.64	0.62	0.66	0.52	0.53	0.58
Goods and service exports	0.50	0.64	0.29	0.66	0.61	0.58	0.62	0.46	0.49	0.54
Foreign direct investment	0.35	0.46	0.14	0.45	0.41	0.46	0.42	0.34	0.37	0.38
Population indicators										
Urban Population	0.50	0.55	0.62	0.58	0.53	0.49	0.59	0.50	0.53	0.55
Military personnel	0.41	0.45	0.58	0.56	0.48	0.41	0.50	0.45	0.44	0.47
Population	0.42	0.45	0.67	0.49	0.45	0.37	0.50	0.44	0.44	0.47
Literate Population	0.40	0.44	0.64	0.49	0.46	0.38	0.48	0.38	0.46	0.46
Technology indicators										
Total PCs	0.53	0.66	0.35	0.69	0.62	0.64	0.64	0.49	0.53	0.57
Internet Users	0.44	0.58	0.24	0.59	0.55	0.53	0.52	0.40	0.44	0.48
Mobile Phones	0.42	0.55	0.26	0.53	0.52	0.50	0.53	0.45	0.42	0.47
Globalization indicators										
Aircraft Departures	0.37	0.52	0.23	0.53	0.50	0.42	0.42	0.35	0.41	0.42
Tourism, arrivals	0.34	0.44	0.13	0.42	0.42	0.42	0.44	0.34	0.35	0.37
Currency transfer from abroad	0.32	0.34	0.24	0.36	0.32	0.40	0.42	0.24	0.32	0.33
Tourism, receipts	0.27	0.44	0.08	0.41	0.37	0.41	0.38	0.30	0.35	0.33
Workers remittances	0.20	0.30	0.27	0.35	0.31	0.48	0.35	0.20	0.23	0.30
Development assistance	0.09	0.07	0.14	0.08	0.08	0.07	0.10	0.10	0.10	0.09
Physical Indicators										
Kilometers of road	0.37	0.41	0.45	0.44	0.41	0.36	0.44	0.36	0.34	0.40
Arable Land	0.28	0.28	0.50	0.29	0.29	0.24	0.31	0.29	0.30	0.31
Surface area	0.21	0.17	0.40	0.21	0.18	0.14	0.23	0.20	0.19	0.22
Freshwater resources	0.07	0.08	0.19	0.14	0.08	0.07	0.09	0.06	0.12	0.10

Three of four economic indicators show strong correlation, though foreign direct investment shows no meaningful correlation. Only one of four population indicators – urban population – shows strong correlation, though the other three show some correlation. Technology indicators fare similarly, with total PCs showing strong correlation and the other two factors showing some correlation.

Six indicators chosen to represent global interconnection correlate poorly to story counts. Aircraft departures shows some correlation, correlating strongly to AltaVista, Google and CNN's results, but modestly overall. No other factors correlate strongly, and none are worse than development aid and assistance, which bears the dubious distinction of least correlated to media attention, a fact that comes as no surprise to anyone who works in the International development community. Four physical indicators – largely reflections of the size of a nation – also fail to correlate meaningfully with article distribution.

It is worth noting how abnormal BBC's results appear in the comparison of nine media sources. Comparing each source's correlation to a given World Bank data set to the average correlation to that data set, BBC is more than one standard deviation away from the norm on 20 of 21 possible indicators. By contrast, three sites are within standard deviation on all 21 indicators, and three other sites are only outside of standard deviation on one or two indicators. CNN is outside the standard deviation on four indicators, and the New York Post is outside on five.

In contrast to all other sites, BBC story distribution shows no meaningful correlation to any economic indicators (it verges on a loose correlation to GDP, with R^2=0.38). It shows strong correlation to all four population indicators, and is the only site to show strong correlation to a physical indicator (Arable land, R^2=0.5)

Why does BBC present such a different statistical profile from other news media outlets? In a word: empire. BBC appears to have an editorial policy that mandates regular coverage of nations formerly in the British Empire. Many of these nations have large populations and small GDPs, and therefore the BBC attention is more closely correlated to population factors than to economic ones. Before anointing the BBC the champion of the poor, it's worth noting that the BBC does not spend noticeably more attention on poor countries in Central Asia or Central America (areas where the British Empire was not colonially involved) than other news media outlets.

It is also interesting to note that the six sites with most similar correlation patters – AP, Reuters, New York Times, Washington Post, AltaVista and Google – represent the shortest timeframes, ranging from 14 to 90 days. It's possible that correlations over a wider timeframe show a different pattern than correlations over a short time period. In other words, if we could examine shorter time slices of CNN and New York Post data, they might look more similar to the data of the six most similar sites.

It remains to be seen whether such correlations will hold true over time. Early results suggest they will be. Correlations performed with Google data on May 5, 2003 – a period that should have no overlapping stories with the period considered in this paper – showed 0.67 correlation to GDP (compared to 0.62 with current data) and 0.44 correlation to population data (compared to 0.45 with current data).

While these correlations and lack of correlations suggest something about media distribution – namely that it may have more to do with economics than with population distribution – they suggest a challenging question: should one really expect media distribution to be connected to these sorts of factors? After all, one reads very few news stories that report that Japan's GDP is still vastly larger than Nigeria's – shouldn't news to be closely correlated to things that occur, like natural disasters and wars?

Fortunately for the world's population, and unfortunately for statisticians, all nations are not uniformly plagued with wars and natural disasters. While it would be statistically convenient to compare the coverage a war in Sudan receives to the coverage a war in

similarly-sized Canada experiences, Canada has been reluctant to comply by engaging in a military conflict.

Instead of working from a 150-data point World Bank set, it is useful to consider Project Ploughshare's Armed Conflict Report[12] , which lists 29 countries "hosting" armed conflicts in 2001, their most recent data set. When one examines CNN results for these 29 nations (because CNN is one of two data sets that includes all of 2001, and the other one, BBC, is not representative of the other eight sets), it becomes clear that that hosting a conflict increases a nation's visibility, but not as much as might be expected. 15 of the 29 have fewer stories than predicted by a population estimation, while eight have more than predicted. (The remaining six are within the predicted range.) The results are almost inverted considering attention versus GDP – 18 of the 29 have more stories than predicted by GDP, while only five have fewer.

	Stories	Pop Estimate	GDP estimate	Pop Variance	GDP Variance
Chad	5	475	110	-98.95 %	-95.50%
Nigeria	623	2959	922	-78.94%	-32.40%
Myanmar	414	1550	1213	-73.30%	-65.87%
Guinea	152	462	166	-67.09%	-8.17%
Senegal	201	545	221	-63.13%	-8.96%
India	5486	11472	4555	-52.18%	20.43%
Burundi	213	436	63	-51.14%	235.67%
Congo Dem. Rep.	815	1634	237	-50.12%	243.45%
Uganda	491	949	252	-48.24%	95.10%
Sudan	631	1177	422	-46.38%	49.48%
Colombia	782	1437	1446	-45.59%	-45.91%
Nepal	534	970	248	-44.95%	115.00%
Algeria	681	1156	1106	-41.08%	-38.41%
Angola	432	674	352	-35.91%	22.84%
Somalia	352	520	204	-32.30%	72.97%
Sierra Leone	363	358	67	1.39%	441.69%
Kenya	1191	1153	397	3.26%	200.09%
Sri Lanka	902	834	494	8.09%	82.74%
Indonesia	4917	4038	2094	21.77%	134.80%
Rwanda	634	476	115	33.23%	453.15%
Turkey	2659	1948	2116	36.49%	25.63%
Iran	2819	1873	1788	50.50%	57.69%
Pakistan	5286	3129	1158	68.95%	356.54%
Philippines	3670	2126	1317	72.64%	178.70%
Russian Federation	10273	3176	3435	223.43%	199.03%
Afghanistan	5866	1066	592	450.23%	891.36%
Yugoslavia	3973	577	385	588.65%	933.02%
Iraq	7796	975	1162	699.98%	570.85%
Israel	7456	412	1728	1709.83%	331.36%

[12] "The Armed Conflict Report 2000", http://www.ploughshares.ca/CONTENT/ACR/ACR00/ACR00.html, accessed July 31,2003.

In other words, a nation hosting a violent conflict appears likely to command more attention than it would simply based on its economic strength. This increased attention is not enough to raise the level of attention to that which a wealthy country of the same size would expect. Sudan demands more attention than Tanzania (similar size, similar size of economy) but less attention than Canada (similar size, much larger economy), despite the fact that Sudan is hosting a violent conflict.

The results listed above are obviously not comprehensive – charts and maps of all results generated by GAP scrapers are available for download at h2odev.law.harvard.edu/ezuckerman[13].

Conclusions

After comparing the GAP profiles of different media outlets, it is possible to make a few broad generalizations:

- Media attention is not homogenous – nations are not covered equally. A small number of nations receive a large share of the attention of a given media outlet.
- No single factor explains the distribution of media attention perfectly. If one estimates distribution based on population figures, fewer stories than expected tend to appear about poor nations and more stories than expected appear about small, wealthy nations. If one estimates based on national GDP, certain large economies, especially in South America, are underrepresented. And, with either estimation, the Middle East is overrepresented.
- While no single factor correlates perfectly to the distribution of media attention, national GDP and imports of goods and services correlate more closely than any other factor. In general, economic and technology factors correlate more closely than population factors. Physical attributes of nations and factors related to international communications and travel do not appear to correlate to media attention distribution.
- Some evidence exists that the relationship between media distribution and population or GDP holds true over both long and short periods of time.
- While six of nine media outlets exhibited very similar behavior, and two others roughly similar behavior, BBC demonstrated radically different patterns. The distribution of BBC's attention is closely correlated to population distribution and not strongly correlated – if at all – to GDP distribution.
- Violent conflict draws attention to a nation, but less than might be expected. A nation hosting a violent conflict will receive more attention than a peaceful nation with a similarly sized economy. It will not receive more attention than a similarly-sized, peaceful nation with a much larger economy, suggesting that GDP may be a more important factor in explaining media distribution than violent conflict.

This paper focuses on correlating factors to observed patterns, rather than trying to demonstrate causality. In particular, the intent of this paper is not to suggest that media

[13] Readers are welcome to download any or all data sets and correlate them to other factors, and this author welcomes correspondence, especially correspondence including additional results.

sources consciously tailor their reporting to national GDP, with editors checking the wealth of nations before deploying reporters abroad.

Figuring out what actually causes media distribution likely requires investigation of entirely different factors. Where do media outlets position their reporters, and how do they make those decisions? How does the ease or difficulty of traveling to a given nation (Myanmar, for instance) influence the amount of attention a media source is able to pay to it? These questions are beyond the scope of this paper, but need to be addressed before suggesting causality of media attention distribution.

A final conclusion of this paper is a warning for all media consumers – *caveat emptor*. It is clear that all news media outlets studied in this paper have large blank spots in their global attention maps. Future GAP papers will attempt to chart these blank spots more accurately and make it possible for media consumers to make better choices or lobby their media outlets for more global coverage.

Future Steps

GAP is intended to be a long-term project, looking at an increasing number of media sources, correlating to a larger universe of data sets and entertaining theories of causality as well as correlation. Some steps likely to be taken in the future:

- **More data sources**. Some of the most interesting future directions for GAP come from comparing similar media sources, like BBC and CNN. It would be ideal to be able to compare the 20 largest newspapers in the US on a regular basis, and to compare US newspapers to international English-language news sources. Some of this is simple legwork – figuring out what certain search engines do and don't support and creating appropriate configuration and keyword files.

  Some of the most interesting news sources are available through information retrieval services like Lexis. Unfortunately, these services are generally configured to be "unscrapeable", using checksums to create custom URLs that could not be requested by automated tools. Fortunately, Lexis does include excellent facilities for performing automated searches and having the results mailed to you. One design for the next scraper takes input from mail, rather than from the web, and relies heavily on mail pre-processing through procmail.

  In the future, GAP will deal with non-English language media as well. This will require a thorough rewrite of keyword lists, but should not require major code changes.

- **Database driven**. Current GAP scripts have no sense of history – they're not aware of what results they generated a week or a month ago. Those analyses need to be performed by hand. In the future, GAP scrapers will log their results into a database, making it possible to see how a particular news source

represented certain search times over a period of time. This is critical for resolving questions about the reliability of data models.

- **Influential media index**. Right now all stories reported by a source like news.google.com have the same weight, whether reported by the Wall Street Journal or the Samoa Observer, despite WSJ's significantly larger reach and influence on the international community. The Influential Media Index will attempt to identify twenty most influential media sources and track their attention on a daily basis, providing both summarized information, and information on how each individual source deviates from the mean.

- **Multiple Factor Correlation.** All correlation studies performed on GAP data thus far consider a single factor at a time. In the next round of analysis, it will be interesting to combine factors and see if any combination gives a near-perfect estimation of media distribution for a particular media source.

- **Newswire Analysis**. Andrew McLaughlin and Diane Cabell, both of the Berkman Center, each independently suggested that GAP could track both large newswires like AP and Reuters, and individual newspapers that use these newswires for foreign coverage. A comparison would reveal a great deal about editorial decisionmaking as concerns international coverage. Do newspapers run very little news on Africa because little is available from wire services? Or do they deprioritize available stories due to perceived lack of reader interest?

- **The Media Coefficient**. One of the most interesting statistics in development economics is the Gini coefficient. It measures the difference between the actual distribution of wealth in a country and the theoretical, perfectly equal distribution. The result is a number between 0 (perfect equality) and 1 (perfect inequality). Using the same technique, media inequality could be computed by comparing the actual story count of a nation to a "perfect" media distribution, where everyone in the world gets equal attention from the media.[14] Alternatively, one could calculate differences between an individual outlet's curve and a mean curve, like the proposed Influential Media Index.

- **Open Source Tools.** In the near future, the scripts behind GAP will be released under GPL or a similar license. It is the author's hope that these tools, primitive as they are, could be useful to other researchers interested in quantitative media analysis. It is further hoped that this paper is the first in a long series of studies, and that the author will not be the only one performing said studies.

[14] Jonathan Zittrain of Harvard Law School points out that the analogy between media and money may well break down. While a world where everyone had the same amount of money might be a very nice place, a world where everyone were equally famous might be very strange. Or might not...

Acknowledgements
Chris Warren wrote the original scraper designed to pull data from Google News – the current ScrapeNews program builds heavily on his code, and this project would not have been possible without his original code and his advice on my code.

Nate Kurz authored Mapper, an extremely useful interface to ImageMagick, which turned the mapping of GAP data from an evening-long manual task to a momentary automatic task. GAP would be far less pretty without his code and his input on my code.

Special thanks to Jesse Ross, server wrangler extraordinare, for his help with getting GAP code to play nicely with the Berkman servers.

GAP relies heavily on two open source tools: Perl and ImageMagick. The existence of tools like these turns GAP from a year-long project to something a reasonably inept programmer can create in a few weeks. Long live Open Source!

Many thanks to colleagues who've read early drafts of this paper and offered critique and advice, especially Gerry Wyckoff, Kira Maginnis, Nate Kurz, Zach Yeskel, Noah Eisenkraft, Andrew McLaughlin, Diane Cabell, and Jonathan Zittrain. Special thanks to Rachel Barenblat and Andrew McLaughlin for their extensive editing.

Thanks to the Berkman Center for Internet and Society at Harvard Law School, and especially director John Palfrey, for supporting this research.