Agenda Setting in Offline and Online Media: Index of Media Freedom

Gleb Romanyuk

May 31, 2011

Abstract

The news topics discussed in traditional offline media like TV, radio and print newspapers differ from ones discussed in online media like blogs. The difference appears to be more evident in countries where traditional media is subject to censorship, heavily influenced by government, and, in short, not absolutely free. At the same time, blogosphere remains relatively free. In this paper, I provide a comparison between the topics discussed in online and offline media and develop an objective, automatically calculated index of media freedom. It appears that the indices based on comparison of online and offline media are positively correlated with survey-based Freedom House indices for the sample of OECD countries and some developing countries. My online-offline media index has an advantage over survey-based indices because it is computationally less costly. A theoretical model is built to illustrate how online media can be less biased than offline media when the government attempts to exert control over news sources in the country.

1 Introduction

The measurement of the level of freedom of the press in a country is of considerable interest to make inferences, to investigate what it depends upon and whether it should be promoted. Several international organizations provide some conventional indices of press freedom. Freedom House’s
Freedom of the Press index measures the level of freedom and editorial independence enjoyed by the press in every nation around the world. Freedom House is registered as a US-based non-government organization, however, receives about 80% of its budget from the U.S. government. The organization is often criticized for furthering the interests of the U.S. government. Reporters Without Borders’ Press Freedom Index is another index of French origin.

These indices are survey-based. The questionnaire is sent to partner organizations as well as to journalists, researchers, jurists and human rights activists. The survey asks questions about direct attacks on journalists and the media as well as other indirect sources of pressure against the free press. Therefore these indices are subjective. However, it is important to have an objective index based on verified variables.

The purpose of the paper is to compare the traditional offline media with online media and to provide an index of media freedom based on the dissimilarities. Offline media refers to print newspapers, magazines, journals, radio, TV, etc. Online media refers to pure online newspapers, blogs, forums, etc. There are a lot of print newspapers that are also published online. They are considered offline since their online image mirrors the offline which are possibly subject to some restrictions.

Gentzkow, Shapiro (2010) investigated the question from the viewpoint of the slant of media outlets in U.S. They constructed an index of media slant that measures whether a news outlet’s language is more similar to a congressional Republican or Democrat. Their index is based on the differences in frequency of certain phrases distinctive for one of the parties (e.g. “death tax” and “estate tax”). They used automated methods to determine partisan phrases based of chi-square statistic. Monroe, Colaresi, and Quinn (2009) propose other techniques for selecting words that capture differences in political speech and for evaluating the relative importance of those words, e.g. Bayesian shrinkage and regularization.

I adopt other approach which captures the dissimilarities between offline and online media in agenda. According to the agenda-setting hypothesis, the media has significant agenda-setting power and therefore influences the people’s attitude towards the current social and political events. Cohen (1963) stated that the press “may not be successful much of the time in telling people what to think,
but it is stunningly successful in telling its readers what to think about. Therefore, the government that has control over the outlets, will want to take advantage of its agenda-setting power to pursue its own aims, for example to change its image for people. Therefore, the government or the strong leader in autocratic regime will want to acquire ownership over media or to have control by other methods. In former case the government will have direct control over the editorial policy. In the latter the government introduces censorship or bribes the media. Ultimately, the government will want to exert control over entire media because, as known from the story with Vladimiro Montesinos, the single station is enough to report all crucial information to the citizens. Offline media is easier to influence and the efficacy is high relative to online media (at least so far). Online media are not so easy to control. As a result, the picture of a day reported by offline and online outlets does not coincide. In extreme cases the difference becomes hard to renounce. For instance, on Dec 24, 2010 the president Medvedev during the interview with the CEOs of main Russian TV channels virtually recognized that there is a “dramatic gap” in agenda between TV and the Internet.

Definitely, there are some inherent differences between offline and online media in agenda setting. For example, online discussions are predisposed to creating memes, in blogs news are not concentrated and it is cheap to publish information online. In my work the politically driven bias in agenda is studied.

Larcinese, Puglisi and Snyder (2008) gauged the extent of agenda bias on economic issues for a large number of U.S. newspapers. They found fairly robust evidence of political partisanship in the coverage of the unemployment rate.

In this paper I focus on the text outlets like newspapers from offline media and blogs from online media. I tried to exclude online newspapers from the analysis since they are somewhat intermediate type which may mix characteristics of pure offline newspapers and blogs.

In this paper I propose a particular method of comparison of the agenda in outlets. I calculate the frequencies public figure is mentioned during particular time period in online and offline resources and then compare them. Distribution of mention frequencies of public figures serves as a proxy to distribution of topics presented in mass media, i.e. the agenda. Every person from every country

---

1 As cited in Larcinese, Puglisi and Snyder (2008)
is associated with two time series which represent the dynamics of the public interest toward this person. Under the free offline media the two time series should be similar. If systematic considerable discrepancies are found it may be inferred that one of media types does not reflect the citizens’ interests and seems to be biased. As blogs are hardly controlled by government, big differences between mentioning tracks are accounted to the distortions in offline media. This method does not perfectly reflect the news picture of the day but it is more robust than working with texts directly and mining the topics from the articles.

The extension of basic method is to discriminate those public figures who have something to do with politics (I will refer to them as politicians) and those who do not. So, I calculate the measures for politicians, for non-politicians and the measure of politicians controlling for non-politicians.

The remainder of the paper is organized as follows. Section 2 presents simple theory on the supply of difference between media types. Section 3 introduces data sources, describes data collection strategy and the data itself. In section 4 I present the indices of media freedom and provide evidence validating these indices against conventional Freedom House’s index. Section 5 concludes.

2 Model of two types of media

The model sheds light on why some types of mass media fall under considerable government influence while others do not. Here I utilize the idea of media capture and the model is based on the model proposed in Besley, Prat (2006) in which I allow for two types of media. I have more detailed model, in which types of mass media have continuous magnitude of bias and where citizens react on the level of bias they perceive in the outlets. It can be found in the appendix. In the model described in this section people’s habits are fixed, that is, the audience for online and offline media are exogenous.

2.1 Setup

Consider a model with three sets of players: media outlets of two types (online and offline), the government and the citizens. It is a two-period retrospective voting model. In the first period an
incumbent is exogenously in power. There are two possible types $\theta \in \{b, g\}$ with $P(\theta = g) = \gamma$, where $g$ stands for “good” and $b$ stands for “bad”. Policy outcome depends solely on the politician’s type. A good incumbent delivers a benefit of one to voters while a bad incumbent generates a payoff of zero. The voters do not observe their payoffs at the time of the reelection decision: the policies may have long-term consequences or voters pay much attention to the social welfare which they fail to observe without media.

There is unlimited pool of potential media outlets, each of which can become active by paying a fixed cost - $c^{on}$ to become an online outlet (to become a blogger) and $c^{off}$ to become an offline outlet. Key assumption is that $c^{on} \ll c^{off}$. The media receives signals which it transmits to the citizens. If the incumbent is good, the media observes no verifiable information. If the incumbent is bad, the media has a chance to receive a signal. The media receives a verifiable signal that the incumbent is bad with probability $q^{on}$ and $q^{off}$ (probability is unconditional on the incumbent’s type). The parameter $q^*$ represents the technological and cultural characteristics and also on institutional variables such as the existence of censorship, the extent of privacy protection regulation, etc. I expect that $q^{on} > q^{off}$ but the idea is that the results of the model holds even if $q^{on} = q^{off}$, so for the rest of the model I assume that there is common parameter $q$. The are three more assumptions. First, signals can only be bad. Second, news cannot be fabricated, so media reports news only if they have the information that the incumbent is bad. Third, all media have the same information.

The media profits from the audience and from transfers from the incumbent. Audience-related revenue may be associated with literate revenue from subscription and advertising or with the mission of reaching as many viewers as possible. The latter seems to be a feature of blogosphere, for example. Viewers prefer informative news. I assume that the audience of each type of media divides itself equally among the media outlets of this type. The audience-related revenue of the outlet is normalized to zero if the outlet has no news, and it is $a^j/m^j$, $j \in \{on, off\}$ if it has news. Parameters $a^{on}$ and $a^{off}$ represents the maximum potential audience-related benefit of online and offline media, respectively; and $m^j$ is the number of outlets of type $j$ that are reporting news. I allow $a^{on}$ to be less that $a^{off}$ but it should not be as small as the costs of entry:
If at least one outlet has informative news, then the entire audience of this type of media become informed. I assume \( qa^j > c^j, \ j \in \{on, off\} \), so that at least one outlet will find it profitable to enter in both media markets.

Incumbent manipulates news. The incumbent makes each outlet \( i \) take-it-or-leave-it offer \( t_i \). A media outlet that accepts this offer suppresses his signal about the politician’s type. Offers are simultaneous and private. A transfer \( t_i \) costs \( t_i \) to the incumbent but yields \( t_i/\tau \) to the media outlet. The parameter \( \tau \) is a transaction cost. The incumbent gets \( r - \sum t_i \) if she is reelected and \(- \sum t_i \) if she is not. The transfers may be interpreted not just as cash bribe but in wider sense as an indirect influence.

The timing of the game is as follows:

1. Each outlet decides if it becomes active or not. If it enters the market, it pays \( c^on \) or \( c^off \).
   Entry decisions are made simultaneously and noncooperatively.

2. The incumbent’s type is realized. If type is good \( \theta = g \), media observe no signal. If \( \theta = b \), media observe bad signal with some probability. The incumbent observes the media signal and selects a transfer \( t_i \geq 0 \) for each outlet \( i \).

3. Media outlet \( i \) observes transfer \( t_i \) and decides to accept or reject it. If it accepts in reports uninformative news and receives \( t_i/\tau \). If it rejects, it reports the true signal and receives audience-related revenue.

4. Audience of each type of media observes the signals reported and vote for the incumbent or a challenger of unknown type. Audience of offline media, which is virtually all citizens, changes the bad incumbent with probability of 1. Audience of online media is limited group of citizens so it changes the incumbent with probability \( \lambda < 1 \). One can think of \( \lambda \) as equal to the ratio of online audience to offline audience \( a^on/a^off \).
In situations where the media receive a transfer in exchange for silence the media is called to be captured, referring otherwise to the media being independent.

### 2.2 Equilibrium

The concept of equilibrium being used is pure-strategy sequential equilibrium.

**Proposition 1.** Denote the maximum number of outlets in independent offline (online) media market by \( m_{\text{off}}^o \) (\( m_{\text{on}}^o \)). Then \( m_{\text{off}}^o = \left\lfloor \frac{\tau a_{\text{off}}}{c_{\text{off}}} \right\rfloor \), \( m_{\text{on}}^o = \left\lfloor \frac{\tau a_{\text{on}}}{c_{\text{on}}} \right\rfloor \).

Equilibrium in the media markets may be one of three kinds:

(a) If \( \frac{1}{r} - \lambda \frac{m_{\text{off}}^o}{\tau a_{\text{off}}} < 1 \), both media are independent. The numbers of active media outlets are \( m_{\text{off}}^f \) in offline media and \( m_{\text{on}}^f \) in online media.

(b) If \( \frac{1}{r} - \lambda \frac{m_{\text{off}}^o}{\tau a_{\text{off}}} > \frac{1}{r} \), offline media is captured, online media is independent. The number of active offline outlets is \( \left\lfloor \frac{(1-\lambda)r}{\tau a_{\text{off}}} \right\rfloor \), the number of active online outlets is \( m_{\text{on}}^f \).

(c) If \( \frac{1}{r} \geq \frac{1}{r} \frac{m_{\text{on}}^o}{\tau a_{\text{on}}} \), both media are captured. The number of active outlets in offline and online media are \( \bar{m} = (m_1, m_2) \) where \( \bar{m} \) is one of the integer maximal points of the set

\[
\left\{(m_1, m_2) : \tau a_{\text{off}} m_1 + \tau a_{\text{on}} m_2 \leq r, m_1 \leq \frac{\lambda r}{\tau a_{\text{on}}}, m_2 \geq m_{\text{off}}^f \right\}.
\]

Third kind of equilibrium is not really expected because \( c_{\text{on}} \) is very small so \( m_{\text{on}}^o = \left\lfloor \frac{\tau a_{\text{on}}}{c_{\text{on}}} \right\rfloor \) is very large. Last condition on \( \bar{m} \) means that the equilibrium numbers of outlets are so that the incumbent has enough rent \( r \) to keep them all silent. Equilibrium numbers \( \bar{m} \) are not uniquely identified because we want to take kind of operation \( \lfloor \cdot \rfloor \) on the set in \( \mathbb{R}^2 \).

If \( r \) is very large, the incumbent can buy both media. The size of \( r \), however, may be substituted by the high costs of entrance on online media market \( c_{\text{on}} \) or low cost of silencing media \( \tau \). If \( r \) is small enough, both media are left independent. If \( r \) is intermediate, online media are left free because of its multiplicity while offline media is captured because of its absolute impact on the voters.
3 Data

3.1 Data Sources

It seems that there is no database ready available on this theme, so I used raw data. In this paper I concentrated on the mention frequency of politicians and popular people. So I have to proceed with two steps.

The first step is to find a list of country’s politicians. For the US, I used Wikipedia’s list of politicians, then added people from some Forbes “most” lists, and simply googled with some appropriate requests. The initial number of persons is 4152. For Russia, I took the Lentapedia list and most popular persons from Yandex’s Press-Portraits service totaling in 1601 persons. For European countries I used European Media Monitor’s (EMM) News Explorer which daily constructs news clusters and extract names (see Appendix for script).

The second step is to generate time series of mention counts. I employed popular search engines, Google and Yandex. I wrote scripts in Perl that send automated requests to the engines and save the returned results for the particular person in the specified time period (see Appendix). For mentionings in offline news Google News Archive Search was used for the US and the EU. For Russia, I used Yandex’s News service, which, in addition, allowed to specify the source of the news, “central offline outlets”.

I investigate the mention dynamics on a year-long time period, from Oct 9, 2009 to Oct 8, 2010. The time step is one week so every time series has 52 observations for the US, Italy, France, Germany, Sweden, Spain, Poland, Turkey, Norway, Portugal and Denmark. For Russia, I have collected mention frequencies for 5-year period with the same time step. As a result, for every country there are two panels. Panel \( \{ n_{ct} \} \) is frequencies of mention in offline news, panel \( \{ b_{ct} \} \) is frequencies of mention in blogs, \( i = 1, \ldots, N \) - index for person, \( t = 1, \ldots, T \) - index for time period, \( c \) - country.

To discriminate politicians and non-politicians I used Google and Yandex as well because lists are large and manual assignment is too burdensome, time consuming and will have vague criteria. For every public figure \( i \) three search queries are sent to the search engine. First contains only
his/her name. Second query demands that the page contain one of the words politician, politics, and policy (and synonyms in other languages than English). Third query demands that the page do not contain neither of the above mentioned words. The count of found results are saved; denote these numbers by $Z_i^0$, $Z_i^+$, and $Z_i^-$, respectively.

3.2 Description of the Data

Firstly, look at the aggregated data. As an illustration, consider two countries: the USA and Russia. At the figure (1a) x axis is popular people from US sorted by $PB(i) = \sum_t b_{it}^U$ - the number of mentionings in blogs for the whole period. Function $\log PB(i)$ is a smooth line in the figure. Noisy function is $\log PN(i) = \log \sum_t n_{it}^U$, log of number of mentionings in offline news. It is normalized by means to be of the same order. Under free media I expect that there would not be great deviations of $\log PN(i)$ from $\log PB(i)$. So the measure of dissimilarities between blogs and newspapers may be based on the size of discrepancies between $PN(i)$ from $PB(i)$. The absolute value of the measure definitely does not give any information but it makes sense if compared with other countries.

Big discrepancies may be analyzed separately. For the US, I label big discrepancies (see figure (1a)) but these outshoots should be accounted to the noisy character of data rather than to dissimilarities in media types since there are no evident “non-free media” story associated with these individuals.

Figure (1b) where all individuals are sorted with respect to $PN(i)$ looks very similar. Moreover, it is again difficult to give intuition for the outshoots.

Figure (2) shows the picture for Russia is similar. It may seem noisier but the number of persons is more than twice as little. As in the case of US, significant outshoots are hardly interpreted as “non-free media” case. So, it is difficult to draw inferences based on such pictures and fully aggregated data.

The surge of interest is reflected in both online media and offline media in quite coordinated fashion. Figure (3) illustrates this correlation. High mention frequency in blogs go along with high mention frequency in newspapers, and the significantly positive relation remains after controlling
(a) Sorted by number of mentionings in blogs

(b) Sorted by number of mentionings in newspapers

Figure 1: People in US blogs and newspapers
Figure 2: People in Russian blogs and newspapers
Collected data \((Z_{i}^{0}, Z_{i}^{+}, Z_{i}^{-})\) allows to make inferences about the extent a public figure has something to do with politics. He or she is more likely to be a politician if the proportion when the name mentioned together with \textit{politician}, \textit{politics}, and \textit{policy} \(Z_{i}^{+}/Z_{i}^{0}\) is great while the proportion when mentioned with neither of these words \(Z_{i}^{-}/Z_{i}^{0}\) is small. Figure [4], panel (a) shows that this line of logic gives the results consistent with common knowledge about the public figures. Denote the likelihood that public figure \(i\) is politician by \(y_{i}\). It can also be interpreted as a degree of being a politician. Ideally, \(Z_{i}^{-}/Z_{i}^{0} = 1 - Z_{i}^{+}/Z_{i}^{0}\) and the second proportion is redundant for deriving \(y_{i}\) but it is not the case in the data due to errors and precision and other limitations of search engine. So I define

\[
y_{i} = \frac{Z_{i}^{+}}{Z_{i}^{0}} - \frac{Z_{i}^{-}}{Z_{i}^{0}}
\]
Figure 4: Politicians and non-politicians, Russia, Yandex data.
4 Measuring Bias

4.1 Indices

Here I describe the measures I will use to compare online and offline data. Given mention frequency of individual $i$ on $t$th week, $\{n^c_{it}\}$ and $\{b^c_{it}\}$, define $p^b_{it}$ and $p^n_{it}$ as relative frequencies of $i$’s mentionings with respect to the other people mentioned on the $t$th week.

\[
p^b_{it} = \frac{b_{it}}{b_t}, \quad b_t = \sum_j b_{jt}
\]

(2)

One can think about $p^b_{it}$ ($p^n_{it}$) as a probability that individual $i$ is mentioned in blogosphere (in newspapers) in $t$th period. Total count of times all public figures are mentioned in blogs in week $t$ is $b_t$ (sum over individuals).

Consider several aggregate indices based on this variables. Large group of indices is defined by common formula

\[
m^2 = \text{med} \sum_{i=1}^{N}(p^b_{it} - p^n_{it})^2 \cdot w_{it}
\]

(3)

that is the squared difference of probability distributions averaged across all listed individuals and all weeks. Number of public figures in the country is $N$. The summands are weighted by $w_{it}$. I take sample median instead of average because the data features frequent outliers.

Remark. $0 \leq \sum_{i=1}^{N}(p^b_{it} - p^n_{it})^2 \leq 2$. It means that indices of the form (3) are bounded and comparable across countries.

I considered a number of weighting functions.

$\textbf{m2a:}$ $w_{it} = 1$

$\textbf{m2b:}$ $w_{it} = p^b_{it} + p^n_{it}$. The rationale behind this and two next indices is to pay more attention to most mentioned public figures and not to get noise from the rarely used ones.

$\textbf{m2b\_blogs:}$ $w_{it} = p^b_{it}$
**m2b_newsp**: $w_{it} = p_{it}^b$

**m2c**: This two indices restrict the number of public figures to $Q$ most mentioned. These indices are auxiliary and are later used to understand whether rarely used public figures has some impact on other indices.

**m2c_blogs**: $m2c\_blogs = \text{med}_t \sum_{i=1}^{Q} (\tilde{p}_{it}^b - \tilde{p}_{it}^n)^2$, $i$ are $Q = 100$ public figures most mentioned in blogs.

**m2c_newsp**: $m2c\_newsp = \text{med}_t \sum_{i=1}^{Q} (\tilde{p}_{it}^b - \tilde{p}_{it}^n)^2$, $i$ are $Q = 100$ public figures most mentioned in newspapers.

**mchi2**: $w_{it} = \frac{1}{\frac{p_{it}^b}{n_t} + \frac{p_{it}^n}{b_t}}$, that is $mchi2 = \text{med}_t \chi^2_t$, where

$$
\chi^2_t = \sum_{i=1}^{N} (p_{it}^b - p_{it}^n)^2 \cdot \frac{1}{\frac{p_{it}^b}{n_t} + \frac{p_{it}^n}{b_t}}
$$

is a $\chi^2$-statistic for testing the hypothesis $H_0 : p_{it}^b = p_{it}^n \quad \forall i$ for two samples from multinominal distribution in week $t$. The idea behind this and three next indices is to apply formal statistical criteria which tests if the distributions match. The more discrepancies between distributions, the greater is the $\chi^2$ statistic.

**mCramer**: $mCramer = \text{med}_t \sqrt{\frac{\chi^2_t}{N_t}}, \quad N_t = b_t + n_t = \sum_j (b_{jt} + n_{jt})$ - total number of mentionings in week $t$. $\phi_c(t) = \sqrt{\frac{\chi^2_t}{N_t}}$ is Cramer’s phi measure - normalization of $\chi^2$ statistic on the number of observations. It is an equivalent for correlation coefficient for two nominal variables, giving a value between 0 and 1.

**mLR**: $mLR = \text{med}_t LR_t$, where

$$
LR_t = -2 \left( \sum b_{it} \ln \frac{p_{it}}{p_{it}^b} + \sum n_{it} \ln \frac{p_{it}}{p_{it}^n} \right)
$$

$p_{it} = \frac{b_{it} + n_{it}}{b_t + n_t}$
Measure is based on LR-test for testing the hypothesis \( H_0 : p_{bi}^t = p_{ni}^t \) \( \forall i \) for two samples from multinomial distribution. \( mLR \) and \( mchi^2 \) are asymptotically equivalent (up to a constant factor) but on the finite sample I use them both.

**mLRn:**

\[
mLRn = \frac{\sum_{t} \sqrt{LR_t}}{N_t}
\]

This attempts to normalize on the size of the samples in the same fashion as Cramer’s phi does. \( N_t = b_t + n_t \) is the sum of lengths of samples.

**mHellinger:**

\[
mHellinger = \frac{\sum_{t} \left( \sqrt{p_{bi}^t} - \sqrt{p_{ni}^t} \right)^2}{\sum_{t} \left( \sqrt{b_{it}} + \sqrt{n_{it}} \right)^2}
\]

This is Hellinger distance between two distributions. Its values lie between 0 and 2. It is asymptotically equivalent to \( \chi^2 \) under the null. See, e.g., Borovkov (1998).

**mDHD:**

\[
mDHD = \frac{\sum_{t} \left( \sqrt{p_{bi}^t} \beta_t + \sqrt{p_{ni}^t} (1 - \beta_t) \right) \sqrt{p_{it}}}{\sum_{t} \left( \sqrt{b_{it}} + \sqrt{n_{it}} \right)^2}, \text{ where}
\]

\[
\beta_t = \frac{b_t}{b_t + n_t}
\]

\[
p_{it} = \frac{(\sqrt{b_{it}} + \sqrt{n_{it}})^2}{\sum_{t} (\sqrt{b_{it}} + \sqrt{n_{it}})^2} = \frac{b_{it} + n_{it} + 2 \sqrt{b_{it} n_{it}}}{b_t + n_t + 2 \sum_t \sqrt{b_{it} n_{it}}} \neq \frac{b_{it} + n_{it}}{b_t + n_t}
\]

This measure is based on the Hellinger deviance test which uses Minimum Hellinger Distances estimators. It is asymptotically equivalent to LR test but is robust to contaminations in some sense (see Van Dijk et al. (1997)).

Figure (5) shows the pairwise scatterplots of indices when data on Russia is collected in Google (label RU_G), figure (6) when in Yandex (label RU_Y). In the rest of the paper I use data from Yandex because it seems to be more full and precise and because “good” measures (see next subsection) give close values of RU_G and RU_Y.
Figure 5: Correlation matrices of indices, data on Russia is from Google.
Figure 6: Correlation matrices of indices, data on Russia is from Yandex
4.2 Sensitivity of the Indices

It is not clear what measure to use in order to get better results. Chi-square and LR statistics have sound theoretical grounds but are not robust to the shocks when \( p_{it} \) is very small and it goes to the denominator. It this subsection I discuss the robustness of the indices derived in the previous subsection to two kinds of shocks. Take two samples of nominal variables with probabilities calculated by \([2]\) - \((p_1, ..., p_N)\) and \((q_1, ..., q_N)\), \(N\) is the number of public figures, \(\sum p_i = \sum q_i = 1\).

Consider two types of shock:

1. The \((N + 1)\)-th public figure is added into the sample and it gets probabilities \(p_{N+1} = \delta\), \(q_{N+1} = 0\), \(\delta\) is small relative to \(p_1, ..., p_N\). The probabilities of public figures 1 to \(N\) are updated to \(\tilde{p}_i = (1 - \delta)p_i\), \(\tilde{q}_i = q_i\), \(i = 1, ... N\). The idea is that the figure with low coverage in media is added, so low that in one type of media (in newspapers, for instance) was not mentioned at all according to the data.

2. The \((N + 1)\)-th public figure is added into the sample and it gets probabilities \(p_{N+1} = \varepsilon\), \(q_{N+1} = \varepsilon\). Old probabilities are updated to \(\tilde{p}_i = (1 - \varepsilon)p_i\), \(\tilde{q}_i = (1 - \varepsilon)q_i\), \(i = 1, ... N\). This shock represents the situation when public figure who has low coverage in both media has equal counts of mentionings in the data.

Analysing these shocks I want to understand if the index is not biased much given that some unimportant figure is not included in the sample. And how the exclusion of the tail of rarely mentioned figures affects the index.

Lemma 1. Consider \(\delta \to 0\) and \(\varepsilon \to 0\). Denote index with shock minus old index by \(\Delta\), ratio of index with shock and old index by \(\frac{\text{new}}{\text{old}}\). Then the asymptotic behavior of indices is as presented in the following table:
<table>
<thead>
<tr>
<th>Statistic</th>
<th>1. ((\delta, 0))</th>
<th>2. ((\varepsilon, \varepsilon))</th>
</tr>
</thead>
<tbody>
<tr>
<td>(m2a \ (w_{it} = 1))</td>
<td>(\Delta = O(\delta)) (\frac{\text{new}}{\text{old}} = (1 - \varepsilon)^2)</td>
<td></td>
</tr>
<tr>
<td>(m2b \ (w_{it} = p_{it}^b + p_{it}^n))</td>
<td>(\Delta = O(\delta)) (\frac{\text{new}}{\text{old}} = (1 - \varepsilon)^3)</td>
<td></td>
</tr>
<tr>
<td>(m\chi2 \ (w_{it} = \frac{1}{\frac{p_{it}^b}{N} + \frac{p_{it}^n}{N}}))</td>
<td>(\Delta = O(\nu))</td>
<td>(\Delta = 0)</td>
</tr>
<tr>
<td>(mLR)</td>
<td>(\Delta = O(\delta))</td>
<td>(\frac{\text{new}}{\text{old}} = (1 - \varepsilon))</td>
</tr>
</tbody>
</table>

where \(\nu\) is the count of extra mentionings that has a proportion \(\delta\) wrt to new total count of mentionings: \(\frac{\nu}{N + \nu} = \delta\).

So the indices have different behavior under these shocks. For instance, the index based on chi-square and LR does not change at all in case of the shock \((\varepsilon, \varepsilon) \\forall \varepsilon\). Under the same shock, the index calculated as the sum of squared differences \((m2a)\) decreases by a factor of \((1 - \varepsilon)^2\). So, the index decreases exponentially when adding many rarely mentioned public figures. Figures (7) and (8) illustrate the point. It is “bad” behavior because it is too determined. However, if the shock of type 2 \((\delta, 0)\) is introduced, the index changes only little \(O(\delta)\). The chi-square under the same \((\delta, 0)\)-shock may change significantly on \(O(1)\), what is comparable to \(\chi^2\) statistic itself, and we see it in the data (see figure (7b)).

Bottom line is that I will prefer \(mLR_n\), \(mCramer\), \(mHellinger\), and \(mDHD\) as candidates for “best” index.
Figure 7: Indices on US data

Figure 8: Indices on Russian and French data
4.3 Comparison with FotP

Calculate indices for the available countries and associate it with Freedom of the Press from Freedom House (FotP). I did four types of regression: (1) index on FotP without controls; (2) controlling on GDP per capita; (3) controlling on index of Internet penetration; (4) controlling on both GDP pc and Internet penetration. The results are in tables [1]-[4], where in panels (a) robust standard errors are used and in panels (b) standard errors are bootstrapped. The slope coefficient of FotP are illustrated in figures [9] and [10].

Results are not robust to the choice of index. Taking robust standard errors, \( m_2a \) and \( m_2b \) have strong correlation with FotP; \( mchi2 \) and \( mLR \) give different sign depending on controls. Indices which prove stable in the sense of Lemma 1 and are correctly normalized (\( mLrn \), \( mCc \), \( mHellinger \), and \( mDHD \)) do not agree with each other: \( mLrn \) and \( mCramer \) are positively correlated with FotP while \( mHellinger \) and \( mDHD \) are negatively correlated. All four associations, however, are insignificant. Bootstrapped s.e. are more appealing in these regressions because the number of observations is very low. In all cases they are much larger than usual standard errors and even for \( m_2a \) and \( m_2b \) the FotP coefficient turned insignificant.

4.4 Indices for Politicians and Non-politicians

Samples of public figures in my dataset contain people of different professions. Although the comparison of pooled distribution of public figures in online media with pooled distribution in offline media allows to find gaps in agenda, distinction between politically related and politically unrelated public figures is important. The agenda about non-politicians (musicians, sportsmen, etc) adds noise to the estimates while differences in their media coverage hardly have something to do with media freedom. So in order to make the analysis more persuasive, I divided all public figures into three groups according to the \( y_i \) defined in (1) - the likelihood that public figure \( i \) is politician. For each country, public figures are sorted in ascending order according to values of \( y_i \). Public figures from the first tercile is then declared to be non-politicians, public figures from the third tercile to be politicians, public figures from the central tercile are declared unidentified and
<table>
<thead>
<tr>
<th>VARIABLES</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
<th>(7)</th>
<th>(8)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>m2a</td>
<td>m2b</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FotP</td>
<td>14.41**</td>
<td>8.988</td>
<td>11.32</td>
<td>8.349</td>
<td>2.023**</td>
<td>2.095***</td>
<td>1.830*</td>
<td>2.042**</td>
</tr>
<tr>
<td></td>
<td>(6.454)</td>
<td>(10.53)</td>
<td>(8.434)</td>
<td>(12.72)</td>
<td>(0.695)</td>
<td>(0.628)</td>
<td>(0.850)</td>
<td>(0.852)</td>
</tr>
<tr>
<td>GDP (PPP) per capita Internet</td>
<td>-0.0131</td>
<td>-0.00814</td>
<td>13.59*</td>
<td>11.80</td>
<td>0.000174</td>
<td>0.000582</td>
<td>0.848</td>
<td>0.976</td>
</tr>
<tr>
<td>penetration %population</td>
<td>(0.0293)</td>
<td>(0.0345)</td>
<td>(6.606)</td>
<td>(10.73)</td>
<td>(0.0191)</td>
<td>(0.00222)</td>
<td>(0.554)</td>
<td>(0.695)</td>
</tr>
<tr>
<td>Observations</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.121</td>
<td>0.138</td>
<td>0.156</td>
<td>0.162</td>
<td>0.220</td>
<td>0.220</td>
<td>0.233</td>
<td>0.235</td>
</tr>
</tbody>
</table>

*** p<0.01, ** p<0.05, * p<0.1
(a) Robust standard errors

<table>
<thead>
<tr>
<th>VARIABLES</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
<th>(7)</th>
<th>(8)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>m2a</td>
<td>m2b</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FotP</td>
<td>14.41</td>
<td>8.988</td>
<td>11.32</td>
<td>8.349</td>
<td>2.023*</td>
<td>2.095</td>
<td>1.830</td>
<td>2.042</td>
</tr>
<tr>
<td></td>
<td>(18.14)</td>
<td>(28.34)</td>
<td>(27.09)</td>
<td>(54.63)</td>
<td>(1.189)</td>
<td>(2.484)</td>
<td>(3.564)</td>
<td>(8.223)</td>
</tr>
<tr>
<td>GDP (PPP) per capita Internet</td>
<td>-0.0131</td>
<td>-0.00814</td>
<td>13.59</td>
<td>11.80</td>
<td>0.000174</td>
<td>0.000582</td>
<td>0.848</td>
<td>0.976</td>
</tr>
<tr>
<td>penetration %population</td>
<td>(0.0341)</td>
<td>(0.0501)</td>
<td>(0.0226)</td>
<td>(0.00473)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>R-squared</td>
<td>1.103</td>
<td>2.795</td>
<td>184.1</td>
<td>615.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*** p<0.01, ** p<0.05, * p<0.1
(b) Bootstrapped standard errors

Table 1: Online-offline indices (m2a and m2b) vs. Freedom of the Press
<table>
<thead>
<tr>
<th>VARIABLES</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
<th>(7)</th>
<th>(8)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FotP</td>
<td>-10.02</td>
<td>21.66</td>
<td>-6.675</td>
<td>21.53</td>
<td>-83.13</td>
<td>783.1</td>
<td>63.51</td>
<td>794.1</td>
</tr>
<tr>
<td></td>
<td>(13.20)</td>
<td>(31.65)</td>
<td>(14.43)</td>
<td>(32.43)</td>
<td>(383.9)</td>
<td>(862.3)</td>
<td>(432.9)</td>
<td>(910.4)</td>
</tr>
<tr>
<td>GDP (PPP) per capita</td>
<td>0.0763</td>
<td>0.0773</td>
<td>2.086</td>
<td>2.001</td>
<td>(0.0934)</td>
<td>(0.106)</td>
<td>(2.579)</td>
<td>(2.946)</td>
</tr>
<tr>
<td>Internet penetration %population</td>
<td>(10.97)</td>
<td>(22.95)</td>
<td>(363.0)</td>
<td>(701.2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.011</td>
<td>0.117</td>
<td>0.018</td>
<td>0.117</td>
<td>0.001</td>
<td>0.091</td>
<td>0.017</td>
<td>0.093</td>
</tr>
</tbody>
</table>

*** p<0.01, ** p<0.05, * p<0.1
(a) Robust s.e

<table>
<thead>
<tr>
<th>VARIABLES</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
<th>(7)</th>
<th>(8)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FotP</td>
<td>-10.02</td>
<td>21.66</td>
<td>-6.675</td>
<td>21.53</td>
<td>-83.13</td>
<td>783.1</td>
<td>63.51</td>
<td>794.1</td>
</tr>
<tr>
<td></td>
<td>(24.76)</td>
<td>(90.68)</td>
<td>(109.7)</td>
<td>(145.2)</td>
<td>(1,458)</td>
<td>(2,398)</td>
<td>(1,897)</td>
<td>(4,500)</td>
</tr>
<tr>
<td>GDP (PPP) per capita</td>
<td>0.0763</td>
<td>0.0773</td>
<td>2.086</td>
<td>2.001</td>
<td>(0.114)</td>
<td>(0.125)</td>
<td>(3.104)</td>
<td>(3.997)</td>
</tr>
<tr>
<td>Internet penetration %population</td>
<td>(2.667)</td>
<td>(9.640)</td>
<td>(83,286)</td>
<td>(257,342)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.011</td>
<td>0.117</td>
<td>0.018</td>
<td>0.117</td>
<td>0.001</td>
<td>0.091</td>
<td>0.017</td>
<td>0.093</td>
</tr>
</tbody>
</table>

*** p<0.01, ** p<0.05, * p<0.1
(b) Bootstrapped s.e

Table 2: Online-offline indices (mchi2 and mLR) vs. Freedom of the Press
<table>
<thead>
<tr>
<th>VARIABLES</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
<th>(7)</th>
<th>(8)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FotP</td>
<td>92.92</td>
<td>277.8</td>
<td>104.0</td>
<td>274.9</td>
<td>111.0</td>
<td>128.7</td>
<td>79.86</td>
<td>120.0</td>
</tr>
<tr>
<td>GDP (PPP) per capita</td>
<td>0.445</td>
<td>0.468</td>
<td>-48.69</td>
<td>54.71</td>
<td>0.0427</td>
<td>0.110</td>
<td>0.839</td>
<td>0.839</td>
</tr>
<tr>
<td>Internet penetration %population</td>
<td>(0.917)</td>
<td>(1.070)</td>
<td></td>
<td>(0.718)</td>
<td></td>
<td>(136.9)</td>
<td></td>
<td>(161.2)</td>
</tr>
<tr>
<td>Observations</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.011</td>
<td>0.056</td>
<td>0.012</td>
<td>0.057</td>
<td>0.027</td>
<td>0.028</td>
<td>0.040</td>
<td>0.045</td>
</tr>
</tbody>
</table>

*** p<0.01, ** p<0.05, * p<0.1

(a) Robust s.e.

<table>
<thead>
<tr>
<th>VARIABLES</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
<th>(7)</th>
<th>(8)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FotP</td>
<td>92.92</td>
<td>277.8</td>
<td>104.0</td>
<td>274.9</td>
<td>111.0</td>
<td>128.7</td>
<td>79.86</td>
<td>120.0</td>
</tr>
<tr>
<td>GDP (PPP) per capita</td>
<td>0.445</td>
<td>0.468</td>
<td>-48.69</td>
<td>54.71</td>
<td>0.0427</td>
<td>0.110</td>
<td>0.839</td>
<td>0.839</td>
</tr>
<tr>
<td>Internet penetration %population</td>
<td>(0.833)</td>
<td>(1.352)</td>
<td></td>
<td>(0.700)</td>
<td></td>
<td>(136.9)</td>
<td></td>
<td>(161.2)</td>
</tr>
<tr>
<td>Observations</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.011</td>
<td>0.056</td>
<td>0.012</td>
<td>0.057</td>
<td>0.027</td>
<td>0.028</td>
<td>0.040</td>
<td>0.045</td>
</tr>
</tbody>
</table>

*** p<0.01, ** p<0.05, * p<0.1

(b) Bootstrapped s.e.

Table 3: Online-offline indices (mLRn and mCramer) vs. Freedom of the Press
<table>
<thead>
<tr>
<th>VARIABLES</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
<th>(7)</th>
<th>(8)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mHellinger</td>
<td>mDHD</td>
<td></td>
<td></td>
<td>mHellinger</td>
<td>mDHD</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FotP</td>
<td>-216.3</td>
<td>-520.3</td>
<td>-308.9*</td>
<td>-535.0</td>
<td>-273.7**</td>
<td>-336.8</td>
<td>-274.3**</td>
<td>-334.9</td>
</tr>
<tr>
<td>(154.7)</td>
<td>(408.5)</td>
<td>(158.3)</td>
<td>(432.8)</td>
<td>(83.46)</td>
<td>(274.1)</td>
<td>(81.81)</td>
<td>(278.1)</td>
<td></td>
</tr>
<tr>
<td>GDP (PPP) per capita</td>
<td>-0.732</td>
<td>-0.619</td>
<td></td>
<td>-0.152</td>
<td>-0.166</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(0.988)</td>
<td>(1.160)</td>
<td>(0.611)</td>
<td></td>
<td>(0.709)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Internet penetration %population</td>
<td>407.2**</td>
<td>270.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(131.2)</td>
<td>(321.9)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.045</td>
<td>0.132</td>
<td>0.096</td>
<td>0.153</td>
<td>0.191</td>
<td>0.201</td>
<td>0.191</td>
<td>0.201</td>
</tr>
</tbody>
</table>

*** p<0.01, ** p<0.05, * p<0.1

(a) Robust s.e.

<table>
<thead>
<tr>
<th>VARIABLES</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
<th>(7)</th>
<th>(8)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mHellinger</td>
<td>mDHD</td>
<td></td>
<td></td>
<td>mHellinger</td>
<td>mDHD</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FotP</td>
<td>-216.3</td>
<td>-520.3</td>
<td>-308.9</td>
<td>-535.0</td>
<td>-273.7</td>
<td>-336.8</td>
<td>-274.3</td>
<td>-334.9</td>
</tr>
<tr>
<td>(490.7)</td>
<td>(599.4)</td>
<td>(607.6)</td>
<td>(1,105)</td>
<td>(258.2)</td>
<td>(341.6)</td>
<td>(352.8)</td>
<td>(606.9)</td>
<td></td>
</tr>
<tr>
<td>GDP (PPP) per capita</td>
<td>-0.732</td>
<td>-0.619</td>
<td></td>
<td>-0.152</td>
<td>-0.166</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(0.934)</td>
<td>(1.434)</td>
<td>(0.603)</td>
<td></td>
<td>(0.948)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Internet penetration %population</td>
<td>407.2</td>
<td>270.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(31,284)</td>
<td>(91,166)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.045</td>
<td>0.132</td>
<td>0.096</td>
<td>0.153</td>
<td>0.191</td>
<td>0.201</td>
<td>0.191</td>
<td>0.201</td>
</tr>
</tbody>
</table>

*** p<0.01, ** p<0.05, * p<0.1

(b) Bootstrapped s.e.

Table 4: Online-offline indices (mHellinger and mDHD) vs. Freedom of the Press
Figure 9: Indices $m2a$, $m2b$, $mchi2$, $mLR$ vs Freedom of the Press (avplots)
Figure 10: Indices mLRn, mCramer, mHellinger, mDHD vs Freedom of the Press (avplots)
I calculate four indices $m_2a$, $m_{Cramer}$, $m_{Hellinger}$, and $m_{DHD}$ separately on the subsample of politicians only and on the subsample of non-politicians only. The association with Freedom of the Press index is presented in figure 11 in the first two rows. The figure shows (first row) that the correlation for politicians is positive for $m_2a$ and $m_{Cramer}$ but remains negative for $m_{Hellinger}$ and $m_{DHD}$. Correlation of FotP and non-politician based indices is mainly positive, moreover, the t-statistics are higher so the correlation is stronger. The third row presents correlation between FotP and politician based indices controlling for index calculated with non-politicians.
5 Conclusion

In this paper, I propose an index of media freedom based on agenda setting in online and offline media. To construct my index I used mention frequencies of popular people in both types of media over the year period. Conventional indices like Freedom of the Press are computed by the big international organizations and demand lots of time and resources to hold of a poll of experts. The approach of analysing the agenda proposes easier way to calculate the index because the calculation process is almost automated. On the cross-country sample, the index is consistent with general ideas about the media freedom in presented countries and for some specification of the index is positively correlated with Freedom House’s index. For other specifications the correlation is negative so the research should be continued because the sample of countries is quite small. Further, I provide a theoretical model which describes why the online media may remains free while the offline media is captured. The model assumes that potential outlets incur much less costs when entering the online media market than the offline media market; the government does not capture online editions because they are very numerous, at the same time having weaker impact on the population.

Another possible way to compare media types is to compare the news topics themselves. Search engines like Google and Yandex clusterize news and sort the clusters by popularity. The clusters therefore represent topics of the day, week, month, etc. So the research may be directed towards the comparison of blogosphere news clusters with offline media news clusters taking into account their temporal change.

References


A Appendix. Alternative model

In this section I propose a more microstructured model which distinguishes the magnitudes of media capture. The concept of media bias is used here. In equilibrium the bias in online media is less that in offline media. This model is based on the model from Gehlbach, Sonin (2009) to which I add some additional features.

A.1 Setup

Consider a model with two sets of players: a continuum of citizens of mass one, indexed by $i \in [0,1]$ and a government that controls media outlets. Media outlets are not state-owned and the
government provide inducements to encourage outlet to bias coverage. The government bear costs for doing that.

At a cost normalized to one, each citizen $i$ may invest in a single project, $\pi_i \in \{0,1\}$. The private return depends on the state of the world $S \in \{L,H\}$, low or high. The government prefers more investment to less regardless of the state, and so it want to mobilize citizens to take actions that are not in their individual best interest. “Mobilization” may be in positive sense, e.g. voting for the incumbent, and in negative sense, e.g. not participating in anti-regime protest.

In state $H$ the project yields a private return $r > 1$, whereas in $L$ the private return is equal to zero. The realized state is not observed by the citizens but they have a common prior belief that the probability of high state is $\theta$. Assume $\theta r < 1$, which implies that in the absence of any additional information citizens do not invest.

There are two sources of news: online media and offline media. For concreteness, I say that citizen “watches TV” or “reads blogs”. Every source of the news contains one of two messages, $\hat{S} \in \{\hat{H}, \hat{L}\}$. Citizens observe these signals and update their priors. Each citizen chooses one of three options: watch TV, read blogs or do not do anything and remain with the prior. Before the realization of the state of the world, the government publicly chooses editorial policies for two types of media $\beta_{TV}(S), \beta_{b}(S) \in [0,1] \times [0,1]$, where $\beta^*(S)$ is the probability that the source reports the message $\hat{H}$ when the state is $S$. Citizen’s posterior that the state of the world is $H$, conditional on having received the message $\hat{S}$, is denoted by $\theta(\hat{S})$. Watching or reading the news may be profitable to citizens if government’s editorial policy $\beta^*(S)$ is sufficiently informative. A citizen, for example, will invest if after receiving the message $\theta(\hat{S}) r > 1$.

Each citizen $i$ has the exogenous costs of watching news $(\mu^TV_i, \mu^b_i)$, where $\mu^TV_i$ and $\mu^b_i$ are random variables distributed independently and uniformly on $[0,m]$. Therefore, there are citizens, for whom watching news is cheaper than reading blogs and vice versa.

The timing of the game is:

1. The government chooses an editorial policy $(\beta^TV(S), \beta^b(S))$, which is observed by all citizens.

2. Each citizen decides if he watches TV, reads blogs or does not do anything.

32
3. The state of the world is realized, with the messages $\hat{S}^{TV}$ and $\hat{S}^{b}$ determined according to $(\beta^{TV}(S), \beta^{b}(S))$. Citizens that chose to use any source of the news receive the messages.

4. Each citizen decides whether to invest in the project.

Government maximizes the citizens’ investment choosing optimal values of $(\beta^{TV}(S), \beta^{b}(S))$. When the government chooses the editorial policy which deviates from the true state of the world, it bears the costs. Here, I do not specify the origin of the costs, yet its linearity is consistent with lobbying mechanism from Gehlbach, Sonin (2009) between government and private media outlet. The government’s utility function is

$$U_{G} = \psi \int_{0}^{1} E_{\pi}i di - \beta^{TV}_{L} - (1 - \beta^{TV}_{H}) - \kappa(\beta^{b}_{L} + 1 - \beta^{b}_{H})$$

(4)

The parameter $\psi$ denotes the strength, with which government values investment. It also means how expensive is the inducements to the media and may be interpreted as the degree of democracy development. The parameter $\kappa > 1$ measures how much inducements into online media is more expensive than into offline media.

A.2 Equilibrium

First notice that it is profitable to invest if the state of the world is high. So in equilibrium when the state of the world is high the media outlets report the truth, that is $\beta^{*}(H) = 1$. In low state of the world they should sometimes also report the truth because otherwise no citizen would choose to watch news.

The posterior probability of having high state of the world conditional on receiving the message $\hat{H}^{j}, j \in \{TV, b\}$:

$$\theta(\hat{H}^{j}) = \frac{\theta}{\theta + (1 - \theta)\beta^{j}_{L}}$$

Expected benefit of watching the news in media type $j$ is
Figure 12: Choice of citizen conditional on common benefits and individual costs

\[ B_j = (\theta + (1 - \theta) \beta^j_L)(\theta(\bar{H}^j)r - 1) = \theta r - \theta - (1 - \theta) \beta^j_L \]  

(5)

Accept preliminary assumption that in equilibrium the bias in blogs is lesser than in TV: \( \beta^b_L < \beta^T_L \). Or, in terms of benefit, \( B^b > B^T \). The expected mass of individuals watching each type of the media may be derived from the payoffs comparison. If citizen reads blogs, he gets \( B^b - \mu^b_i \), if he watches TV, he gets \( B^T - \mu^T_i \), otherwise he gets zero. Figure (12) plots the areas of best choices of citizens.

The expected mass of individuals who watch TV is

\[ P(TV) = P(B^T > \mu^T_i, B^T - \mu^T_i > B^b - \mu^b_i) = \frac{m - B^b + \frac{B^T}{2}}{m^2} B^T \]

The expected mass of individuals who read blogs is

\[ P(blogs) = P(B^b > \mu^b_i, B^T - \mu^T_i < B^b - \mu^b_i) = \frac{m B^b - \frac{(B^T)^2}{2}}{m^2} \]

The probability that a citizen invests conditional on watching \( j \) is equal to \( \theta + (1 - \theta) \beta^j_L \). And
so the expression for the government’s utility (4) may be rewritten as

\[ U_G = P(TV)(\theta + (1 - \theta)\beta^TV_L) + P(blogs)(\theta + (1 - \theta)\beta^b_L) - \beta^TV_L - \kappa \beta^b_L \]

The variables \( \beta^j_H \) do not enter into the expression because they are previously shown to equal one in equilibrium.

**Proposition.** *Equilibrium level of bias in blogs is less than in TV: \( \beta^b_L < \beta^TV_L^* \).*

**Proof.** I abuse notation to simplify the argument and from this moment assume that citizens’ benefits \((B^b, B^{TV})\) enter the expression for \( U_G \) instead of \((\beta^b_L, \beta^TV_L)\). This is justified because according to (5) \( B^* \) is a linear function of \( \beta^L_\cdot \). Rewrite the government’s maximization problem in terms of the benefits:

\[
U_G = \frac{1}{m^2} \left[ m\theta r (B^b + B^{TV}) - m \left( (B^b)^2 + (B^{TV})^2 \right) - B^b B^{TV} \theta r \right] + \frac{1}{2m^2} \left( B^{TV} \right)^2 (3B^b - B^{TV}) + B^{TV} + \kappa B^b \rightarrow \max_{B^b, B^{TV}}
\]

Analytical solution is not beautiful because FOCs are quadratic polynomials but function \( U_G \) may be analyzed to compare the equilibrium values of \( B^b^* \) and \( B^{TV^*} \). Note that the first part of \( U_G \) (in square brackets) is symmetric so it cannot generate unequal \( B^b^* \) and \( B^{TV^*} \). The second part is skewed to the area where \( B^b > B^{TV} \) because \( \kappa > 1 \). So, maximizing (6) on the bounded area, the government chooses \( B^b^* > B^{TV^*} \). This means that \( \beta^b_L^* < \beta^TV_L^* \), the bias in online media is less than in offline media.

\[ \square \]

**B Appendix. Listing of scripts**

`gblogs.pl - acquires data from Google Blogs Search`

`#!/usr/bin/perl`
use strict;
use feature 'say';
use Data::Dumper;
use HTML::Tree;
use LWP::Simple;
use Encode;
use POSIX qw/mktime/;

# # # # # # # # # # # # # # # SUBS # # # # # # # # # # # # # # # # # # # # # # # # # #
sub makeUrl {
    my ($text, $d0, $m0, $y0, $d1, $m1, $y1, $l) = @_; 
    my @parts = ();
    push @parts, "http://blogsearch.google.com/blogsearch?";
    push @parts, "as_q=" . $text;  # ordinary search
    push @parts, "as_mind=" . $d0;
    push @parts, "as_minm=" . $m0;
    push @parts, "as_miny=" . $y0;
    push @parts, "as_maxd=" . $d1;
    push @parts, "as_maxm=" . $m1;
    push @parts, "as_maxy=" . $y1;
    push @parts, "lr=lang_" . $l;
    push @parts, "num=10";
    push @parts, "hl=en&scoring=d&c2coff=1&safe=off&ie=UTF-8&ctz=−180&c2coff=1&btnG=Search+Blog&as_oq=&as_epq=&as_eq=";

    return join('&', @parts);
}

sub getFreq {

}
my $url = shift;
my $html=undef;
my $att=0;
do {
    $html = get($url);
} while (not defined $html and
    ++$att and
    sleep 1 and
    say STDERR "\a$att attempts failed! $url") ;

$html =~ /of about\D+\[(\d+,)+\] \D+ for/; return 0 unless defined $1;
#
my $n=$1;
$n =~ s/\//g; # udalit zapyatuyu kak razdelitel razryadov
return $n;
#

# # # # # # # # # # # # # BEGIN # # # # # # # # # # # # # # # # # # # # # #
my $lang = pop @ARGV; # the 1st argument
die unless defined $lang;
#
my %data = () ;
$|++;
for my $p (@people) {
    $p =~ s/\s+/+/g;
    my $counter = 0;
    for my $i (@periods_seq) {
        my @from = (localtime($today−$period∗($i+1))[3..5]; # +1, chtob peri
                      v
        sleyuyuuii den
        my @to = (localtime($today−$period∗($i−1)))[3..5];
        $from[1]++; $to[1]++;
        my $url = makeUrl($p,@from,@to,$lang);
        #push @urls_to_send, [$url,$p,$i−1];
        $data{$p}[$i−1] = getFreq($url);
        print STDERR "$p	period/uni2423$i/uni2423from/uni2423$nperiods	" . $data{$p}[$i−1] . "\n";
        if (++$counter == 4) {
            if ($data{$p}[$periods_seq[0]−1] == 0 and
                   $data{$p}[$periods_seq[1]−1] == 0 and
                   $data{$p}[$periods_seq[2]−1] == 0 and
                   $data{$p}[$periods_seq[3]−1] == 0) {
                say STDERR "Seems_to_be_all_zeros..Next_person...";
                map { $data{$p}[$_] = 0 } 0..($nperiods−1);
                last;
            }
        }
        #usleep(100000); # microseconds, 0.1 seconds
        #sleep 2;
    }
   say join("	", $p, @{$data{$p}});
}
$|−−;

gnews.pl - acquires data from Google News Archive Search

#!/usr/bin/perl −w

use strict;
use feature 'say';
use Data::Dumper;
use LWP::Simple;
use POSIX qw/mktime ceil/;

# SUBS

sub makeUrl
{
    my ($text, $d0, $m0, $y0, $d1, $m1, $y1, $l) = @_; 
    ($d0, $m0, $d1, $m1) = map { sprintf("%02d",$_) } $d0, $m0, $d1, $m1;
    my @parts = ();
    my $domain = $l;
    if ($l eq 'en' || $l eq 'tr') { $domain = 'com'; }
    return "http://news.google.$domain/archivesearch?as_q=$text&num=10&hl=en&btnG=Search+Archives&as_epq=&as_oq=&as_eq=&as_user_ldate=$m0/$d0/$y0&as_user_hdate=$m1/$d1/$y1&lr=lang_$l&as_src=&as_price=p0&as_scoring=a";
}

sub genBrowser
{
    my @browsers = qw(Firefox Opera Chrome);
    my $b = int(rand(3));
    my @ver = (int(rand(10)), int(rand(20)), int(rand(20)));
    return $browsers[$b] . "/" . join('.', @ver);
}

sub getFreq
{
    my $url = shift;
    my $att=0;
    my $html;

    do {
        $html = get($url);
    }
if ( ($html and $html =~ /but it appears your computer is sending automated requests/) or $att >= 3) {
    say STDERR "IP_banned,\$url,Going_to_sleep\a";
    $html = undef;
    sleep 120;
    say STDERR "Wake_up!";
    $att=0;
} while ( not $html and ++$att and
        sleep 1 and
        say STDERR "\a\$att_attempts_failed,\$url\a");
$html =~ /of about\D{\([\d,]+\)\D{\d+} for}/i;
return 0 unless defined $1;
my $n=$1;
$n =~ s//g; # u dalit zapyatuyu kak razdelitel razryadov
return $n;

BEGIN
my $lang = pop @ARGV; # the 1st argument
die unless defined $lang;
my $period = 7*24*60*60; # in secs - a week
my $nperiods = 52; # a year
my @periods_seq = (1, $nperiods, ceil($nperiods/2), ceil($nperiods/3), 2..($nperiods -1));
my $today = mktime(59,59,23,8,9,110); # 8 Oct 2010

my @people = ();
@people = <>;
chomp @people;

my %data = ();
$|++;
for my $p (@people) {
    $p =~ s/\s+/+/g;
    my $counter = 0;
    for my $i (@periods_seq) {
        my @from = (localtime($today−$period*$i+1))[3..5];
        my @to = (localtime($today−$period*($i−1)))[3..5];
        $from[1]++; $to[1]++;
        my $url = makeUrl($p, @from, @to, $lang);
        #push @urls_to_send, [url, $p, $i−1];
        $data{$p}{$i−1} = getFreq($url);
        print STDERR "$p	period/uni2423$i/uni2423from/uni2423$nperiods\t" . $data{$p}{$i−1} . '\n';
        if (++$counter == 4) {
            if ($data{$p}{$periods_seq[0]−1} == 0 and
                $data{$p}{$periods_seq[1]−1} == 0 and
                $data{$p}{$periods_seq[2]−1} == 0 and
                $data{$p}{$periods_seq[3]−1} == 0) {
                say STDERR "Seems to be all zeros. Next person...";
                map { $data{$p}{$_}=0 } 0..($nperiods−1);
                last;
            }
        }
    }
    #usleep(100000); # microseconds, 0.1 seconds
    #sleep 2;
} say join("\t" , $p, @{$data{$p}} );
}$|--;