Representation of Political Discussions in Web Forums: A Cross-National Assessment

Hai Liang[1] and Fei Shen[2]

1. Web Mining Lab, Dept. of Media & Communication, City University of Hong Kong
2. Dept. of Media & Communication, City University of Hong Kong

Abstract

Gauging public opinion through user generated content (UGC) on social media has experienced an explosive growth in recent years. Although social media have been celebrated for the equality of public expression and large participants involved, the representativeness of online opinions was called into question. This study demonstrates that public expression on the internet is unequally distributed across issues and the interests of the vocal minority and silent majority exhibit a substantial discrepancy. Through a cross national analysis of web-based political discussion forums from 54 societies with 1,218,698 threads, this study found that the user generated content in web forums is socially constructed. The inequality in reply and view distribution and discrepancy between lurker and participants are structured by political system, culture values, and so on. All these findings suggest that online user generated content as another symbolic representation of reality cannot represent the general public opinion or even the opinions of general internet users. Furthermore, the social construction of political discussion on the internet indicates what measured through UGC and the survey results are two different things in nature.

Keywords:

Public opinion, internet representation, social construction of reality, cross national comparison, internet forum, political discussion, lurker

Introduction

There are a growing number of papers using online user generated content (UGC) (e.g., Gonzalez-Bailon, Banchs, & Kaltenbrunner, 2012; Livne, Simmons, Adar, & Adamic, 2011; O'Connor, Balasubramanyan, Routledge, & Smith, 2010; Tumasjan, Sprenger, Sandner, & Welpe, 2010) and search query data (e.g., Granka, 2010; Ripberger, 2011; Scharkow & Vogelgesang, 2011) as a measure of public opinion in recent years. A basic assumption in these studies is that there are more and more people who are accessible to the internet and social media in particular for public expression. To somehow, internet users can represent the general population. However, the representativeness of online discussions faces both empirical and theoretical challenges:

First, access to the internet is not distributed equally. Not every age, gender, race, social group is equally represented on the internet (e.g., DiMaggio, Hargittai, Neuman, & Robinson, 2001). Second, the self-selection bias is commonly observed in online political communication. The user generated content is produced by those politically active (e.g., Himelboim, 2008, 2011; Himelboim, Gleave, & Smith, 2009; Mustafaraj, Finn, Whitlock, & Metaxas, 2011). The silent majority is a huge problem. Users might be reluctant to publicize their opinions (Albrecht, 2006; Jones, 1997). And this lurking behavior may make the gauge of public opinion through UGC biased towards the activists' orientation (Mustafaraj et al., 2011). Third, in addition to the empirical concerns on representativeness of online public opinion, the UGC based opinions might be theoretically different from the results from random sampling survey. It is possible that the presentation of public opinion on the Internet might be socially constructed and structured

The annual conference of the World Association for Public Opinion Research, Hong Kong, June 14-16, 2012.
Correspondence should be addressed to rainfireliang@gmail.com.

by political, cultural, and economic environment where political discussion embedded. The principle of one-person one-vote axiom, which the poll methodology relies on, is apparently violated in forum discussions and hence the UGC based measures might be called into question.

The present study focuses on the second and third challenges to show how users' participations and attentions are distributed in web-based political discussions; whether the vocal minority and silent majority share similar interests; and further to demonstrate how societal-level factors can influence the equality of public expression and discrepancy between lurkers' and participations' interests across discussion topics in web forums. We argue that online user generated content as another symbolic representation of reality cannot represent the general public opinion or even the opinions of general internet users. Furthermore, the social construction of political discussion on the internet indicates what measured from social media and the polling results are two different things in nature.

Literature Review

Representation and Representativeness of Online Political Discussion

Due to the growing number of individuals who were accessible to the Internet around the world, optimistic scholars argue that internet technologies have the potential to make politics more inclusive by provide information and unrestricted communication (e.g., Mitra, 2001; Papacharissi, 2002). Representation of public opinion on the Internet was expected to have the characteristics of diversity, equality, unbiased, and un-restrictiveness (Himelboim, 2011). Based on this premise, researchers began to propose alternative barometers to gauge public opinion through opinionated texts generated in social media and search quires. Although most of papers suggest that this method meets a certain level of face validity when compared with random sampling polls, several problems remain in question.

One of the first is so-called "digital divide" (DiMaggio et al., 2001). Access to the internet is not distributed equally but follow well know factors of inequality, such as gender, age, education, internet skill, and so on. Unequal access to the internet implies that certain groups of people are overrepresented on the internet, whereas others are underrepresented. Samples are impossible to be representative by gathering online expressions (e.g., young, white, and highly educated) (Albrecht, 2006). Nevertheless, it also could be consistent between online and offline public opinion in this situation. First, the distributions of contrast positions could be parallel across different demographics. Second, researchers can weight online opinions according to demographic variables to adjust online opinions to the offline (Gayo-Avello, 2012).

Second, previous studies validated online data by simply comparing aggregated data with corresponding survey results. Correlation between them suggests a certain level of validity. However, due to the lack of survey datasets, issues been selected in comparisons were not randomly sampled but by convenience. Therefore, the correlations might only exist in the top discussed issues, such as presidential approval rating (e.g., Gonzalez-Bailon et al., 2012; O'Connor et al., 2010; Tumasjan et al., 2010), health care, global warming, terrorism in US (e.g., Ripberger, 2011). It is hardly to estimate the correlation between online and offline opinion on non-popular issues unless we can get the population of issues in a society.

Third, people's attention and participation in online political discussions are definitely not a random process as polling methodology assumed. Opinion expression in online political discussion is a process of self-selection: the active seeking of liked or interested contents, or avoidance of contradicted or disliked views (Mutz & Young, 2011). This self-selection bias makes the popularity unequally distributed across issues (Barabasi & Albert, 1999). That means most of the issues introduced on the internet receive insufficient amount of attention and depth of discussion. Furthermore, there could be a significant discrepancy between the distributions of participation and attention due to the discrepant interests of lurkers and participants. Therefore, measuring public opinion through the representation of political discussion on the internet must be biased from that of general population and general internet users.

If representativeness is defined in the same way as public opinion poll, the first two problems can be solved by sophisticated techniques by weighting and gathering issue population. However, the third problem implies what measured from online political discussions is not the same thing as public opinion collected through polling. The polling mythology is heavily based on the "one-person one-vote" principle: every person is equally considered in democracies (Dahl, 1989). Yet, the unequal distributions of attention and participation directly challenged this assumption. Previous studies have shown that some inequalities were replicated in online discussion forums at the individual level. For example, political discussions in newsgroups are hierarchical with a small number of participants who received most of the replies (e.g., Fisher, Smith, & Welser, 2006; Himelboim, 2008, 2011; Himelboim et al., 2009). And the privileged segments of the population were overrepresented in political forums (Davis, 1999; Hill & Hughes, 1998; Wilhelm, 2000). Our argument is beyond this individual-level inequity of internet use, because the individual inequality faces the similar challenges to the digital divide problem. Self-selection bias is a more serious problem: only a small number of political issues proposed on the internet can invoke attention and participation. Therefore, the self-selection process at the individual level, which is reflected in the inequality of popularity across issues, makes the representation of political discussions on the internet biased.

The Social Construction of Online Political Discussion

In addition to show the unequal distribution beyond the individual level, the present study will go further to demonstrate this unequal representation of political discussion in web forums is shaped by political and cultural factors. On the one hand, the dependency on societal factors makes the online public opinion can hardly be consistent with that collected through random surveys; on the other hand, it suggest that online representation of political discussion reflects the institutional and cultural variances across societies. Therefore, we should not downgrade the meanings of online opinions.

It has long been know that representation of social reality on traditional media is socially constructed. The production of media image is not neutral but evinces the power and point of view of the political and economic elites (Gamson, Croteau, Hoynes, & Sasson, 1992). Various factors have been discussed in the process of media representation of reality: organization of news (Tuchman, 1978), source and routine channels (Gans, 1979; Sigal, 1973), ownership of media market (Bagdikian, 1990), media attention cycles (McCarthy, McPhail, & Smith, 1996), and so on. These studies suggest that media representation is not the mirror reflection of the real world at all. Therefore, it is not valid to gauge public opinion through media reports.

Internet technologies were expected to break the restrictions placed upon media representation by societies, politics, and markets (Bagdikian, 2004; Papacharissi, 2002, 2004). Due to the scarce of space, traditional media need to select only a portion of the information individuals need. As mention above, the selection process is not neutral but restricted by many external factors. The internet, on the other hand, provides infinite space where people suggest topics for discussions. The media-generated image of the world has been substituted by user-generated image on the internet. Although, this transformation changes the preconditions for unrestricted political discussion on the internet, it does not mean discussions on the internet are free of any structural constrain. Chinn and Fairlie (2007) found that global digital divide (internet penetration rate) is significantly associated with economic development and quality of regulation. Beyond the level of access to the internet, the second source of bias is how users actually use the platforms to meet their social ends. Although internet technologies provide countless channels for public expression, attention – not information – becomes the scarce source (Nye, 2002). Himelboim (2008) further found that the attentions to various topics on newsgroup are distributed very unequal. However, he did not mention any social factors shaping this pattern.

Internet is just one of the new media technologies, when it provides preconditions for equal and unrestrictive expression; it is embedded in a boarder political, cultural, and economic space. Liberating features of new technologies are not deterministic. New technologies will be molded to fit traditional politics, adapt to current political culture (McChesney, 1996) and structured by real life social relations (Fernback, 1997). Online political discussion is a form of symbolic representation, which is similar to what the critical interpretation of opinion polls (see, for example, Bourdieu, 1979; Herbst, 1993; Salmon and Glasser, 1995; Lewis, 1999; 2001), but structured in a different way. The critical literature on opinion polling criticizes that opinion surveys are ideologically controlled by the pollsters who established the framework and sets the parameters. In online political discussions, the structural constrains might exert influence not through a pollster but through the public's self-selection bias process. In other words, the self-selection bias is conditional on external contextual factors. Internet users are embedded in a broad political and cultural environment; their behaviors on the internet are constrained by these political and cultural characteristics. We are arguing that these characteristics can significantly influence the representation of political discussion on the internet by influencing the degree of self-selection bias.

Political Discussion in Web Forums

Online discussion platform has long been celebrated as a democratizing technology (e.g., Corrado & Firestone, 1996; Rheingold, 1993). It allows citizens to introduce and participate in diverse public issues. Unlike traditional Usenet newsgroups, the web-based forums are not based on e-mail list groups. Discussions in web forums are always organized in discussion threads. A thread is a collection of a seed post and replies, usually display from the oldest to latest. People can freely initiate conversations by posting a seed post and reply to others' replies within a thread. When people browse numerous discussion threads in web forums, they are unrestricted to decide whether to post a reply or not. Therefore, the advantage of choosing web forums to Usenet groups is we know exactly the proportion of lurkers in a thread. Furthermore, since the current study is interested in the representation of political discussion, the issues (rather than participants) been discussed in web forums are the main focus. The threaded discussion also makes the unit of analysis go beyond the individual level.

A thread can be considered as a public issue or political agenda which the seed poster wants to discuss with others. Seed posts in discussion threads play an important role in shaping topic agenda (Himelboim et al., 2009) and opinion expression (Yun & Park, 2011). When plenty of information is accessible in online discussion groups, attention becomes the scarce resource (Nye, 2002). People are free to join discussions on the Internet and suggest new topics for discussion. It could be the case that the availability of a growing number of sources leads to a narrowing of the scope of news and views to which people choose to expose themselves (Sunstein, 2001). Participants preferred joining the already-popular discussion threads due to the limited ability of users to perceive and process a large amount of information. Most of the issues presented in newsgroups evoked little or no discussion and most likely less attention (Himelboim, 2011). Individuals' self-selection behaviors will leads to the "preferential attachment" phenomenon: the rich get richer (Barabasi & Albert, 1999). In web discussion forums, the popular threads will get more popular, therefore,

Hypothesis 1a: The number of replies received by each thread is distributed unequally.

Previous studies measure attention by the number of replies to new threads which is more likely to be a measure of participation (e.g., Himelboim, 2008, 2011; Himelboim et al., 2009). In the present study, we differentiate the two concepts between participation and attention. Web forums explicitly provide the data of number of views in addition to number of replies; therefore we can estimate attention allocation more accurately using the information of views. The unequal allocation of attention suggests the equality and diversity that voices or channels do not necessarily represent the diversity of content presented in discussions. As Jones (1997) suggested that the internet allows us to "shout more loudly, but whether other fellows listen, beyond the few individuals who may reply, or occasional lurker is questionable" (p.30).

Hypothesis 1b: The number of views received by each thread is distributed unequally.

There are many lurkers in web forums who didn't engage in the discussion when viewing the threads. These lurkers are definitely interested in the issues suggested in the threads, but didn't post a reply for social psychological reasons (Yun & Park, 2011). Previous study suggests that the behaviors between lurkers and participants show significant differences (Mustafaraj et al., 2011). Active users in web forums do not represent the general users who are interested in the specific issues.

Hypothesis 2a: There is a substantial discrepancy between the lurkers and participation in terms of issue interest.

It implies that the reply distribution and view distribution are different. Theoretically, attention is a prerequisite for participation. Therefore, the attention allocation process should be less constrained by following the popular than participation process. If the reply distribution is more unequally distributed than view distribution, then the inequality in participation cannot be attributed to the difference of interest in political issues, but to external factors.

Hypothesis 2b: The view distribution is less skewed than the reply distribution.

Web forums users are embedded in social environment rather than in isolation. What generated in web forums should, to some extent, reflect this environment. However, previous studies on political participation and discussion are dominantly focused on the individual predictors with little attention to broader social factors. This study extends the social-economic perspective on internet use to the social construction perspective. The inequalities in content representations are created by users' self-selection posting behaviors. As long as the social environment can influence online user behaviors, it will show impact on the user generated content as well.

Political system could be one of the structural constrains. Political participation can have different meanings in different political systems (Xie & Jaeger, 2008). In democratic societies, political participation on the internet is a way to enhance the degree and quality of public participation in government (Kakabadse, Kakabadse, & Kouzmin, 2003; Noveck, 2003). However, in more restrictive societies, restrictive policies and regulations on online political discussion are made to protect political dominance (e.g., Tan, Mueller, & Foster, 1997). By a cross national comparison, S. Verba, Nie, and Kim (1987)found that institutional system can equalize political activities across social groups in electoral political systems.

Research Question 1: Whether and how does political system influence the equalities in reply and view distributions and discrepancy between lurkers and participants in online political discussions?

National culture may be another structural constrains on online political discussions. Culture is an underlying framework, consisting of the objective reality as manifested in societal institutions and the subjective reality which comprise socialized predispositions and beliefs that guides individuals' perceptions of observed events and personal interactions, and the selection of appropriate responses in social situations (Johansson, 1997). For example, people in individualistic societies may be less likely to follow the popular, thus the reply and view distribution should be more equal than the collectivist societies. However, there are studies show that computer mediated communication has attenuated the social-psychological influences on public expression (e.g., Ho & McLeod, 2008; Yun & Park, 2011). Internet based discussion might be attenuate the cultural impacts.

Research Question 2: Whether and how do cultural factors influence the equalities in reply and view distributions and discrepancy between lurkers and participants in online political discussions?

Method

Data Collection

The data for this study contains two parts: discussion forum data and country-level predictor data. The discussion forum data was collected in three steps.

First, a list of 262 countries was obtained from Internet World Stats, a worldwide internet statistics organization. Second, a series of Google search was conducted through using the keyword "forum & politics & country name." Not all countries have their own discussion forums. As a matter of fact, for a large percentage of countries, we did not find any. When multiple numbers of discussion forums were found for a country, the most popular one will be used – popularity being defined as having the largest amount of posts in "politics" section of the forums. A total of 54 countries or territories and their corresponding forums were identified (see Appendix for details). The list covers a diverse range of countries speaking 18 languages from Asia (20), North America (3), South America (5), Europe (15), Africa (8), and Oceania (3). Our list contains 26 out of the top 40 largest economies in the world. The list also has small economies such as Kyrgzstan, Zimbabwe, and Trinidad and Tobago. Third, all threads in the "politics" sections of the selected forums were downloaded for analysis. The crawling process spanned from September 2011 to March 2012. We used Easy Web Extract, a web scraping software for the crawling task. Most discussion forums use commonly available database management systems (e.g., Dizcuz!, vBulletin, etc) which are highly similar in terms of their structures. Each section of a forum contains a table which tabulates all posted threads. Each thread will be given a unique *URL* address. We scraped all content from the *URL* addresses of the threads from "politics" sections of the selected forums. For each thread, the following information was retained: *URL* of the thread, title of the thread, time and dates of the thread, content of the thread, number of views (i.e., the number of internet users who clicked on the thread to read its content), number of replies the thread received (i.e., the number of internet users who offered their comments), and authors' screen names. A total of 1,218,698 threads were captured. The societal-level predictor data were from different secondary sources. We focused on five aspects: cultural characteristics, modernity, political system, the economy, and internet penetration. Details of these indicators will be elaborated in the ensuing section.

Measure

Inequalities in participation and attention are measured by Gini coefficients of the reply and view distributions. The Gini coefficient measures the inequality among values of a frequency distribution. A Gini coefficient of zero expresses perfect equality where all values are the same. A Gini coefficient of one expresses maximal inequality among values. The frequency distribution of the number of replies received by each thread describes the distribution of participation in each discussion topic. The frequency distribution of the number of views received by each thread describes the distribution of attention allocation across different discussion topics. We calculated the Gini coefficients in each society separately and overall.

Inequality in Reply/view ratio. Reply/view ratio for a thread is the key endogenous variable in the current study. We took the ratio of the number of views and the number of replies a thread received to quantify the percentage of internet surfers who expressed their viewpoints after reading a thread. Conceptually, it speaks to the level of willingness to engage in political discussions and dialogues. The purpose of using a ratio measure here is to control for topic popularity. It could be interpreted as the participation likelihood when attention to each thread is equal. Similarly, we use Gini coefficient of the reply/view ratio distribution as a measure of inequality. It indicates the inequality of the likelihood of participation in online political discussions when control the users' interest differences.

Discrepancy of interests between lurkers and participants indicates whether the interest of peripheral participants correlates with the interest of active participants by calculating whether the posts with a higher number of views also generate a higher number of replies. The findings are indicated by a lurking interest index chart (see Wu, 2008). Figure 1 describes the distribution of all the replies among views (the relation between the accumulative percentage of replies and the total lead postings ranked by the number of views). Lurking interest index equals Gini coefficient (the closer the Gini coefficient is toward 1, the more it indicates a positive correlation). If the interest distribution is even (meaning the number of replies is related more negatively with the number of views), the value of the Gini coefficient is 0. If the interest distribution is highly skewed, which means that the number of replies correlates more positively to the number of hits, the value would go toward 1. Thus the discrepancy index is calculated by 1-the lurking interest index.

Figure 1 about here

National culture dimensions. One of the most comprehensive quantitative studies on cultural difference comes from Hofstede's research on dimensions of national culture (Hofstede, 2001; Hofstede, Hofstede, & Minkov, 2010). We included four Hofstede's cultural dimensions into our study: power distance (PDI), individualism versus collectivism (IDV), masculinity versus femininity (MAS), and uncertainty avoidance (UAI). Data on these four dimensions were harvested from Hofstede's official website[1]. All measures are on a 01-120 scale. Society with a high PDI score tends to accept a hierarchical order and no justification is need for such inequality; society with a high IDV score prefers the notion that individuals are responsible for themselves in a loosely-knit social framework; society with a high MAS score emphasizes achievement and assertiveness more than cooperation and modesty; society with a high UAI score exhibit low levels of toleration toward future uncertainty and ambiguity. In our sample, the average scores for PDI, IDV, MAS, and UAI are 62.6 (*SD*=22.2), 42.1 (*SD*=24.2), 54.0 (*SD*=12.7), 65.3 (*SD*=22.4).

Value orientation. In addition to national culture dimensions, two important value orientation indicators were included in the study: traditional/secular-rational values, and survival/self-expression values (Ingelhart & Welzel, 2005). Unlike Hofstede's cultural indicators, the two value orientation indicators help distinguish traditional societies from modernized secular societies. The data for the two value orientation dimensions were obtained from the World Value Survey (WVS) website[2]. As of today, five waves of WVS have been conducted. We used summary statistics from the most recent wave of the survey fielded in 2006 (if the scores are not available in 2006, we use the most recent wave). Higher traditional/secular-rational value score means more traditional secular-rational value orientation whereas higher survival/self-expression value score means emphasis of self-expression and quality of life.

Political system. There are quite a few publicly available scoring systems aiming to characterize the political systems of the countries across the global (e.g., Freedom House's Freedom in the World index[3]; the Democracy Index compiled by the Economist Intelligence Unit[4], etc.). For this study, we chose Marshall and Jaggers' Polity IV scheme[5]. We picked this scheme for a couple of reasons. First, the Polity scheme is a refined measure which examines "concomitant qualities of democratic and autocratic authority in governing institutions, rather than discreet and mutually exclusive forms of governance." The Polity Score uses a 21-point scale ranging from -10 (hereditary monarchy) to +10 (consolidated democracy). Second, the Polity IV dataset covers all major states in the global system and monitors annual changes. Third, the Polity Score was the most widely used data sources in political science research (Ringen, 2011). The mean of Polity IV score of our sample is 5.5 (*SD*=6.0). The sample for this study included consolidated democracy such as Germany, incoherent authority regimes such as Singapore, and autocracies such as Saudi Arabia.

Control variables. Three control variables were included: GDP per capita, internet penetration, and the total number of threads from the selected forums.

[1] http://geert-hofstede.com/index.php

[2] http://www.worldvaluessurvey.org/wvs/articles/folder_published/article_base_54

[3] http://www.freedomhouse.org/report-types/freedom-world

[4] http://www.eiu.com/public/thankyou_download.aspx?activity=download&campaignid=DemocracyIndex2011

[5] The "Polity IV Project: Political Regime Characteristics and Transitions, 1800-2010" was sponsored by the Political Instability Task Force, which is funded by the Central Intelligence Agency of the US Government.

Data on GDP per capita was obtained from The World Bank[6] (*M*= 16,916 Current US dollar, *SD*= 15,812). Internet penetration data in 2011 was collected from Internet World Stats[7] (M= 9.0, *SD*=1.3). The country with the highest penetration rate included in our study is Australia (.898), but there are other relatively less developed countries, for instance, Ghana (.08), Guinea (.02), and Cote D'Ivoire (0.045). Total number of threads of the selected forums (*M*=22,568 *SD* =41,554) varies from 1,559 (Russia) to 270,462 (China). These three variables were controlled in our analysis because of their potential direct or indirect impacts on people's willingness to participate in online political discussions. First, the economy could be related to people's passion for political engagement in that economic performance and democratic status are highly correlated. Second, internet penetration matters because in areas with low penetration, users are mostly social elites, and their online behavior might be different from grassroots users in a country with high internet penetration. Third, the popularity of discussion forums varies across the globe. In countries where information industry is more developed, for instance, the US, people might prefer using Facebook or Twitter over traditional discussion forums. When few people use this tool, the general enthusiasm toward political discussion through forums could be attenuated.

Result

Unequal Participation and Attention

Findings suggest that the number of replies and views are unequally distributed as a whole. *Figure2* shows that few threads attract a disproportional number of replies and views. However, it is not a classic power-law distribution as previous studies found. Gini coefficient is a more accurate measure of inequality than the power-law coefficient -1.64. Gini coefficient of reply distribution is .94 with 95% confidence interval [.90, .98] which indicates a highly unequal distribution. It suggests that the participation in online political discussions is concentrated on few threads. Overall, 18.7% threads didn't receive any replies. Only 975 (.08%) threads received more than 100 replies. The mean of the number of replies is 129 (*SD* =40209), the median is 8. *Hypothesis 1a* is confirmed. The distribution of the number of views received by each thread is unequal too. The Gini coefficient is .80 with 95% confidence interval [.79, .81]. On average, each thread received 2833 views (*SD* =54247), the median is 563. *Hypothesis 1b* is conformed.

Both the reply distribution and view distribution are highly skewed. As we hypothesized (*hypothesis 2b*), the Gini coefficient of view distribution is much smaller than the Gini coefficient of reply distribution (difference =.14 is significant at .01 level). Figure 2 visually presents the significant difference between reply and view distributions. Attention allocation is distributed more equally than participation in political discussions. Readers should note that we also found a significant correlation between the number of replies and the number of views (.71, p <.001). We use the rely/view ratio as an adjusted measure of the participation. Result shows that even control the attention inequality, the Gini coefficient of the participation likelihood is .63.

In summary, the unequal distributions of participation and allocation of attention suggest that equality and diversity that voices or channels do not necessarily represent the diversity of content presented in discussions. Since most of the threads cannot successfully arouse sufficient discussions, gauging public opinion through the disproportional replies is usually unrepresentative in terms of polling methodology.

Figure 2 about here

Discrepancy of Interests

There are a lot of lurkers in online political discussions. We found that only 2% of users actually expressed their ideas when they were viewing the threads (98% lurkers). *Hypothesis 2a* states that there is a substantial discrepancy between the lurkers and participation in terms of issue interest. The discrepancy index is .34 as a whole (p < .001). It suggests a low level but significant discrepancy between lurkers' interests and participants' interests. The most discrepant society is Hungary .99, whereas

[6] http://data.worldbank.org/indicator/NY.GDP.PCAP.CD

[7] http://www.internetworldstats.com/stats.htm

the least discrepant society is Taiwan, the index is nearly 0. *Figure 3* shows that the inequalities in reply distribution in the 54 societies are significantly different from the inequalities in view distribution ($t = 7.554$, $p < .001$). The result further confirmed *hypothesis 2b*: the reply and view distribution are different. And it also indicates participation pattern is different from the attention allocation pattern in online political discussions. The discrepancy between lurkers and participants' interests further demonstrates that the disproportional participations in different political discussions do not merely reflect users' difference of political interests. Gauging public opinion through replies can neither match opinion polls nor internet users' general interests.

Figure 3 about here

Social Construction of Political Discussion

Table 1 and Figure 2 show that there are cross-society differences in terms of inequalities and discrepancy. RQ1 and RQ2 stated that the inequalities and discrepancy could be explained by external social factors. *Table2* presents that societal level factors are associated with the inequalities and discrepancy in various ways. For the inequality of reply distribution, the degree is positively correlated with self-expression value in society. In a society people emphasis more on self-expression and quality of life, the self-selection is more likely to happen and result in more unequal distribution of replies. The inequality of views is also positively associated with culture values. The more rational and self-expressed society exhibits more focused participation and attention. In addition, cultural tradition also shows significant impacts. We use the reply/view ratio to exclude the relationship between the number of replies and views, result shows that the inequality of participation likelihood is negatively associated with degree of democracy. The more democratic a society is, the more equal people are likely to post replies when controlling people's interest in specific topic. That means the unequal representation of public opinion in web forums is structured by political system in additional to users' interest. On the other hand, attention pattern is more likely to be structured by cultural factors. Furthermore, the last column in table2 shows that the discrepancy between participation and attention is negatively associated with culture values and positively associated with internet penetration. In rational and self-expressed society, the discrepancy is smaller than that in traditional and survival-oriented society. If there are more internet users in a society, the discrepancy is larger than the counterparts. The models fit the data well, R squares range from 30.0% to 43.2%.

Table 2 about here

Discussion

The current study falsified two premises to gauging public opinion through user generated contents on the internet: the equality of participation in online political participation and a mirror reflection of inequality of users' interests. Instead, we conclude that opinions on the internet are social representations structured by political systems and cultural values. First, we presented the participation and attention in political discussions are distributed in a highly skewed manner. Only a small number of threads received many replies and views and many threads received few of them. It could be reasonable to estimate public opinion when topics can successfully evoke sufficient number of replies, but for most of topics it is difficult to estimate public opinion un-biasedly. Moreover, the pretty unequal distribution also implies a process of self-selection. People are likely to reply to or view the threads which were popular. This process violates the random sampling process required by polling methodology.

Second, it is possible that online public opinion could be a mirror reflection of people's interest differences. People replied to few popular threads just because of the individuals' difference of personal interests in different topics. In this sense, it appears that online public opinion might be more accurate than random survey results. However, it is not the case. There is a significant discrepancy between participation and attention. In most of the societies, participants' interests are different from the lurkers'. Therefore, gauging public opinion through UGC cannot even represent the general opinion of internet users not mention to the real world.

Finally, our results suggest a social construction explanation of internet representation of political discussion. The inequalities and discrepancy are influenced by various social factors in various ways. Besides the self-selection process and the discrepancy between participation and attention, what measured on the internet and the polling results are different things in nature. It is theoretically impossible that variability of polling results in different societies varies with political system, culture values, and culture traditions. According to normative democracy theories (e.g., Habermas, [1962] 1989), access to political debate must open for any person affected by the issue at stake. In this study, we found that the users who were interested in the topics were not proportional to express their ideas in the threads. And the proportion of expression and discrepancy between lurkers and participants are significantly associated with the type of political system and culture values etc.

Although, the internet has change the way of representation from media generated reports to user generated content, it didn't change the social essence of internet as one of the new media technologies. What presented on the internet was shaped by political and cultural systems. Normatively speaking, it is far from the ideal representation of political discussion as democracy theories argued. Empirically speaking, it is far from what public opinion polls claimed: "sample surveys provide the closest approximation to an unbiased representation of the public because participation in a survey requires no resources and because surveys eliminate the selection bias inherent in the fact that participation in politics are self-selected" (Sidney Verba, 1996, p. 3).

The Internet, at least, web discussion forum, is another channel to create symbolic realities. We demonstrated these symbolic expressions of reality are shaped by political and cultural structures. Although, the online representation of political discussions cannot represent the general public as defined on the "one-person, one-vote" principal, it does reflect the offline structures of societies. People will mold the internet to fit traditional politics (Hill & Hughes, 1998) and structured by offline social relations (Fernback, 1997). Furthermore, the symbolic representation of political discussions might exert influences on the subsequent expression as stated in the theories of media effect (e.g., Price, Nir, & Cappella, 2006; Yun & Park, 2011). Thus, if public opinion is defined as collective force which works in political process (Crespi, 1997), the "biased" representation on the internet could be more predictable than the "unbiased" ones.

References

Albrecht, S. (2006). Whose voice is heard in online deliberation?: A study of participation and representation in political debates on the internet. *Information, Communication & Society, 9*(1), 62-82. doi: 10.1080/13691180500519548

Bagdikian, B. (1990). *The media monopoly* (3rd ed.). Boston: Beacon.

Bagdikian, B. (2004). *The new media monopoly*. Boston: Beacon.

Barabasi, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science, 286*(5439), 509-512.

Chinn, M. D., & Fairlie, R. W. (2007). The determinants of the global digital divide: a cross-country analysis of computer and internet penetration. *Oxford Economic Papers-New Series, 59*(1), 16-44. doi: Doi 10.1093/Oep/Gpl024

Corrado, A., & Firestone, C. M. (Eds.). (1996). *Elections in cyberspace: Toward a new era in American politics*. Washington, DC: Aspen Institute.

Crespi, I. (1997). *The public opinion process: How the people speak*. Mahwah, NJ: Erlbaum.

Dahl, R. (1989). *Democracy and its Critics*. New Haven: Yale University Press.

Davis, R. (1999). *The Web of Politics: The Internet's Impact on the American Political System*. New York: Oxford University Press.

DiMaggio, P., Hargittai, E., Neuman, W. R., & Robinson, J. P. (2001). Social implications of the Internet. *Annual Review of Sociology, 27*, 307-336.

Fernback, J. (1997). The individual within the collective: Virtual ideology and the realization of collective principles. In S. G. Jones (Ed.), *Virtual* *culture: Identity and communication in cybersociety* (pp. 36-54). Thousand Oaks, CA: Sage.

Fisher, D., Smith, M., & Welser, H. T. (2006). *You are who you talk to: Detecting roles in Usenet newsgroups.* Paper presented at the the 39th Hawaii International Conference on System Sciences, Hawaii.

Gamson, W. A., Croteau, D., Hoynes, W., & Sasson, T. (1992). Media Images and the Social Construction of Reality. *Annual Review of Sociology, 18,* 373-393.

Gans, H. (1979). *Deciding What's News*. New York: Random House.

Gayo-Avello, D. (2012). "I Wanted to Predict Elections with {Twitter} and all I got was this Lousy Paper" - A Balanced Survey on Election Prediction using {Twitter} Data. doi: citeulike-article-id:10623742

Gonzalez-Bailon, S., Banchs, R. E., & Kaltenbrunner, A. (2012). Emotions, Public Opinion, and US Presidential Approval Rates: A 5-Year Analysis of Online Political Discussions. *Human Communication Research, 38*(2). doi: DOI 10.1111/j.1468-2958.2011.01423.x

Granka, L. A. (2010). The Politics of Search: A Decade Retrospective. *Information Society, 26*(5), 364-374. doi: Doi 10.1080/01972243.2010.511560

Habermas, J. ([1962] 1989). *The structural transformation of the public sphere*. Cambridge, MA: MIT Press.

Hill, K. A., & Hughes, J. E. (1998). *Cyberpolitics: Citizen activism in the age of the Internet*. Lanham, MD: Rowman & Little.

Himelboim, I. (2008). Reply distribution in online discussions: A comparative network analysis of political and health newsgroups. *Journal of Computer-Mediated Communication, 14*(1), 156-177. doi: DOI 10.1111/j.1083-6101.2008.01435.x

Himelboim, I. (2011). Civil Society and Online Political Discourse: The Network Structure of Unrestricted Discussions. *Communication Research, 38*(5), 634-659. doi: Doi 10.1177/0093650210384853

Himelboim, I., Gleave, E., & Smith, M. (2009). Discussion catalysts in online political discussions: Content importers and conversation starters. *Journal of Computer-Mediated Communication, 14*(4), 771-789. doi: DOI 10.1111/j.1083-6101.2009.01470.x

Ho, S. S., & McLeod, D. M. (2008). Social-psychological influences on opinion expression in face-to-face and computer-mediated communication. [Article]. *Communication Research, 35*(2), 190-207. doi: 10.1177/0093650207313159

Hofstede, G. (2001). *Culture's consequences: Comparing values, behaviors, institutions and organization across nations*. Thousand Oaks CA: Sage Publications.

Hofstede, G., Hofstede, G. J., & Minkov, M. (2010). *Cultures and organizations: Software of the mind* (3rd ed.). USA: McGraw-Hill

Ingelhart, R., & Welzel, C. (2005). *Modernization, cultural change and democracy*. New York: Cambridge University Press.

Johansson, J. K. (1997). *Global Marketing*. New York, NY: MCGraw-Hill.

Jones, S. G. (1997). The Internet and its Social Landscape. In S. G. Jones (Ed.), *Virtual Culture: Identity and Communication in Cybersociety* (pp. 7-35). Thousand Oaks, CA: Sage.

Kakabadse, A., Kakabadse, N. K., & Kouzmin, A. (2003). Reinventing the democratic governance project through information technology? A growing agenda for debate. *Public Administration Review, 63*(1), 44-60.

Livne, A., Simmons, M. P., Adar, E., & Adamic, L. A. (2011). *The Party is Over Here: Structure and Content in the 2010 Election.* Paper presented at the Fifth International AAAI Conference on Weblogs and Social Media, Barcelona, Spain.

McCarthy, J. D., McPhail, C., & Smith, J. (1996). Images of protest: Dimensions of selection bias in media coverage of Washington demonstrations, 1982 and 1991. *American Sociological Review, 61*(3), 478-499.

McChesney, R. W. (1996). The Internet and US communication policy-making in historical and critical perspective. *Journal of Communication, 46*(1), 98-124.

Mitra, A. (2001). Marginal voices in cyberspace. *New Media & Society, 3*(1), 29-48.

Mustafaraj, E., Finn, S., Whitlock, C., & Metaxas, P. (2011). Vocal minority versus silent majority:

Discovering the opionions of the long tail. *Proc. of IEEE SocialCom*.

Mutz, D. C., & Young, L. (2011). Communication and Public Opinion. *Public Opinion Quarterly, 75*(5), 1018-1044. doi: Doi 10.1093/Poq/Nfr052

Noveck, B. S. (2003). Designing deliberative democracy in cyberspace: The role of the cyber-lawyer. *Boston University Journal of Science and Technology, 9*, 1-91.

Nye, J. S., Jr. (2002). *The paradox of American power: Why the world's superpower can't go it alone*. Oxford, UK: Oxford University Press.

O'Connor, B., Balasubramanyan, R., Routledge, B. R., & Smith, N. A. (2010). *From tweets to polls: Linking text sentiment to public opinion time series.* Paper presented at the ICWSM-2010.

Papacharissi, Z. (2002). The virtual sphere: The internet as a public sphere. *New Media & Society, 4*(1), 9-27.

Papacharissi, Z. (2004). Democracy online: Civility, politeness, and the democratic potential of online political discussion groups. *New Media & Society, 6*, 259-283.

Price, V., Nir, L., & Cappella, J. N. (2006). Normative and informational influences in online political discussions. [Article]. *Communication Theory, 16*(1), 47-74. doi: 10.1111/j.1468-2885.2006.00005.x

Rheingold, H. (1993). *The virtual community: Homesteading on the electronic frontier*. Reading, MA: Addison-Wesley.

Ringen, S. (2011). The measurement of democracy: Towards a new paradigm. *Society, 48*(1), 12-16.

Ripberger, J. T. (2011). Capturing Curiosity: Using Internet Search Trends to Measure Public Attentiveness. *Policy Studies Journal, 39*(2), 239-259. doi: DOI 10.1111/j.1541-0072.2011.00406.x

Scharkow, M., & Vogelgesang, J. (2011). Measuring the public agenda using search engine queries. *International Journal of Public Opinion Research, 23*(1), 104-113. doi: Doi 10.1093/Ijpor/Edq048

Sigal, L. V. (1973). *Reports and officials*. Lexington, Mass: Health.

Sunstein, C. R. (2001). *Republic.com*. Princeton,NJ: Princeton University Press.

Tan, Z., Mueller, M., & Foster, W. (1997). China's new Internet regulations: Two steps forward, one step back. *Communication of the ACM, 40*, 11-16.

Tuchman, G. (1978). *Making News*. New York: Free Press.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Election forecasts with Twitter: How 140 characters reflect the political landscape. *Social Science Computer Review*.

Verba, S. (1996). The citizen as respondent: Sample surveys and American democracy. *American Political Science Review 90*, 1-7.

Verba, S., Nie, N. H., & Kim, J. (1987). *Particioation and political equality: A seven-nation comparison*. Chicago and London: The University of Chicago Press.

Wilhelm, A. G. (2000). *Democracy in the digital age: Challenges to political life in cyberspace*. London: Routledge.

Wu, M. (2008). Measuring political debate on the Chinese Internet forum. *Javnost-the Public, 15*(2), 93-110.

Xie, B., & Jaeger, P. T. (2008). Older adults and political participation on the Internet: a cross-cultural comparison of the USA and China. [Comparative Study Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S.]. *J Cross Cult Gerontol, 23*(1), 1-15. doi: 10.1007/s10823-007-9050-6

Yun, G. W., & Park, S.-Y. (2011). Selective Posting: Willingness to post a message online. *Journal of Computer-Mediated Communication, 16*(2), 201-227. doi: 10.1111/j.1083-6101.2010.01533.x

Table 1 Descriptive Statistics of the Main measures at the Society Level

Measure	*Mean*	*SD*	*Min*	*Max*
Inequality of reply distribution	.706	.100	.450	.997
Inequality of view distribution	.578	.128	.301	.955
Inequality of reply to view ratio	.473	.155	.205	.928
Discrepancy index	.379	.139	.000	.990

Table 2 The Coefficients (standard error) for Robust OLS Regression Model to Predict the Variability of Public Expression in Different Societies

IV	Ineq. of replies	Ineq. of views	Ineq. of Prob.	Discrepancy
Political institution				
Democracy	-.009	.011	-.026*	.000
	(.006)	(.008)	(.011)	(.005)
Cross culture value				
Traditional-Rational	.066	.177**	-.060	-.099**
	(.040)	(.063)	(.089)	(.035)
Survival-Self expression	.082*	.152*	-.025	-.116*
	(.039)	(.067)	(.095)	(.048)
Cultural tradition				
PDI	.002	.004*	.000	-.001
	(.001)	(.001)	(.003)	(.001)
IDV	-.002	-.006*	.003	.003
	(.002)	(.002)	(.003)	(.002)
MAS	.000	.004	-.001	.001
	(.002)	(.003)	(.005)	(.002)
UAI	.000	-.004*	.004	.002*
	(.000)	(.002)	(.002)	(.001)
Control variable				
GDP per capital (log)	.010	-.035	.028	-.005
	(.019)	(.036)	(.075)	(.026)
Internet penetration	-0.242	-.163	-.050	.349*
	(.123)	(.311)		(.147)
Number of threads	.024	-.022	.086	-.004
	(.018)	(0.33)	(.041)	(.021)
Intercept	-.448	.097	-2.025*	.031
	(.224)	(.513)	(.849)	(.328)
R^2	32.2%	44.1%	30.0%	43.2%
N	40	40	40	40

**. Significance at the 0.01 level

*. Significance at the 0.05 level

Figure 1: Lurking and participants interest index. Gini Coefficient = B/(A+B), Discrepancy Index = A/(A+B)=1-Gini Coefficient.

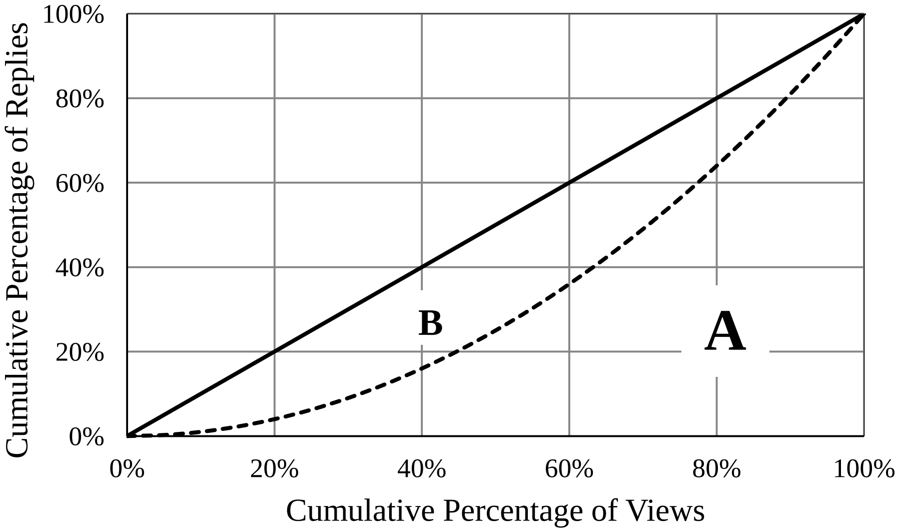


Figure 2: Distributions of the number of replies and views received by each thread

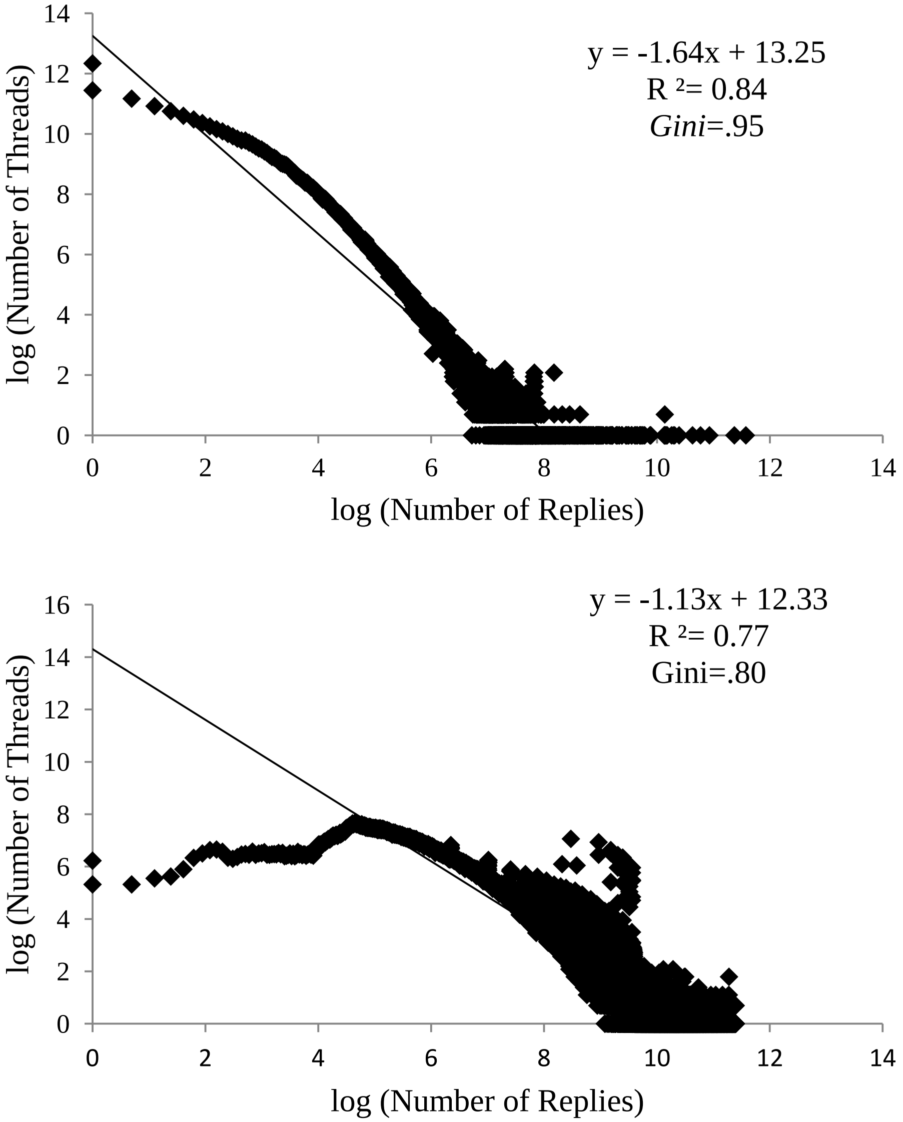
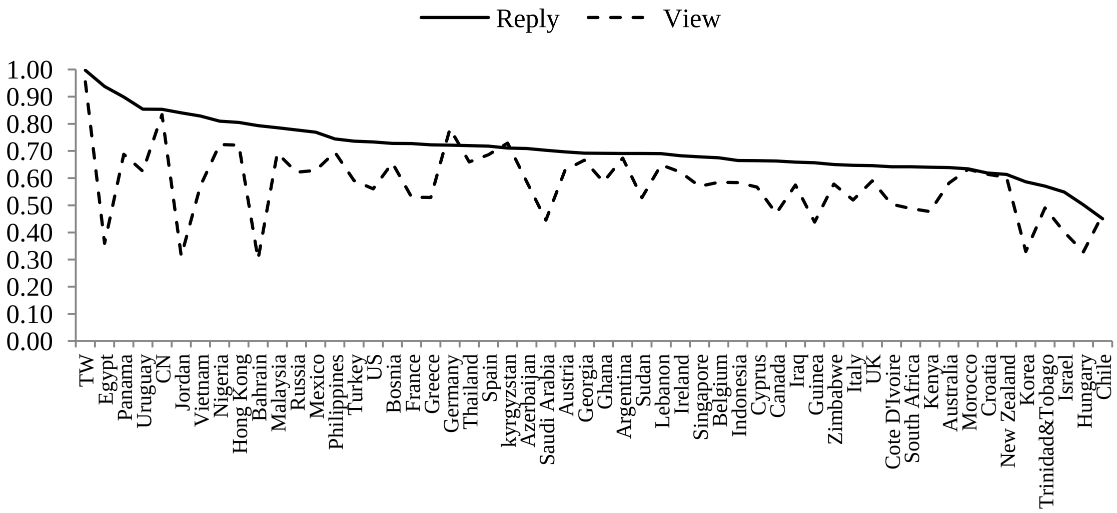


Figure 3: Inequalities in reply and view distributions across 54 societies



Appendix 1

Selected Forums, Societies, URL, Language, and Number of Threads

Region	Forum URL	Language	Number of threads
Argentina	*www.elforro.com*	Spanish	3,755
Australia	*www.ozpolitic.com*	English	3,932
Austria	*www.esoterikforum.at*	German	3,011
Belgium	*forum.politics.be*	Dutch	11,151
Canada	*www.mapleleafweb.com*	English	15,726
Chile	*www.antronio.com*	Spanish	6,078
China	*club.kdnet.net*	Chinese	270,462
Croatia	*www.forum.hr*	Croatian	11,532
Egypt	*forum.egypt.com*	Arabic	33,685
France	*www.forumfr.com*	French	41,261
Germany	*forum.piratenpartei.de*	German	6,321
Ghana	*discussions.ghanaweb.com*	English	5,228
Greece	*www.forums.gr*	Greek	3,364
Hungary	*forum.index.hu*	Hungarian	80,931
Indonesia	*www.indoforum.org*	Indonesian	4,719
Iraq	*www.dijlh.net*	Arabic	8,553
Ireland	*www.politics.ie*	English	11,834
Israel	*www.elsf.net*	Hebrew	11,779
Italy	*forum.kataweb.it*	Italian	29,451
Jordan	*www.amman-stock.com*	Arabic	3,041
Korea	*dvdprime.donga.com*	Korean	39,436
Malaysia	*forum.cari.com.my*	English	15,635
Mexico	*foro.univision.com*	Spanish	9,843
Morocco	*www.wladbladi.net*	French	8,870
New Zealand	*www.gpforums.co.nz*	English	3,749
Nigeria	*www.nairaland.com*	English	61,442
Philippines	*www.istorya.net*	English	5,567
Russia	*forum.baikal.net*	Russian	1,559
Saudi Arabia	*www.aldees.net*	Arabic	3,438
Singapore	*sgforums.com*	English	13,229
South Africa	*mybroadband.co.za*	English	22,609
Spain	*www.elforo.com*	Spanish	12,435
Thailand	*www.thaivisa.com*	English	18,060
Trinidad & Tobago	*www.ttonline.org*	English	3,567
Turkey	*www.siyasiforum.net*	Turkish	9,600
Taiwan	*www06.eyny.com*	Chinese	13,639
United Kingdom	*www.vote-2007.co.uk*	English	4,255
Uruguay	*candombeando.mundoforo.com*	Spanish	13,604
United States	*www.usmessageboard.com*	English	96,025
Vietnam	*v1.ydan.org*	Vietnamese	4,176
Total	*N=40*	17	926,552

Note: Only the 40 societies with complete data are listed here