

Politics, religion and culture in the Internet – structural differences

Albert Hupa<sup>1</sup>

Institute of Applied Social Sciences, Warsaw University

Key words

topic, discourse, topology, crawling, community

---

<sup>1</sup> Albert Hupa, The Institute of Applied Social Sciences, Warsaw University, 69 Nowy Świat St., 00-046 Warsaw, Poland, mail: [albert.hupa@gmail.com](mailto:albert.hupa@gmail.com)

This paper is about the structure of Internet communities, understood as the collection of websites – nodes in the networks heavily connected by hyperlinks and sharing the same topic, i.e. purposeful web graphs (Ackland 2005, 5). People interacting about a given theme create interconnected web pages and by such means they share information, quarrel, discuss, agree or disagree, associate themselves with others or not. The structure of hyperlinks between such pages may provide an insight into the nature of topics which bind such communities.

In order to present this case I open this paper with the presentation of a few approaches to the issue of a community and then I follow on to define discourse and interaction on the Internet. In the following point I describe data mining and sketch possibilities of application of various measures to describe a discourse. Next, on the example of themes of politics, Christianity and culture in the Polish Internet there are presented three different discourse structures. In the end the paper is outlined by conclusions.

## 1 Web communities

Community is one of the most ambiguous notions in social sciences, especially when it comes to the Internet analysis. Robert K. Merton suggested a definition of a group, saying that „the sociological concept of a group refers to a number of people who interact with one another in accord with established patterns” (1957, 286). Merton's definition is much closer to the notion of virtual or cyber community made famous by Howard Rheingold. He described it as „social aggregations that emerge from the Net when enough people carry on those public discussions long enough, with sufficient human feeling, to form webs of personal relationships in cyberspace” (1998, introduction). The case of community looks different on the grounds of hyperlinks topology. Basing on a cluster structure, Gary Flake, Steve Lawrence and Lee Giles define a community on the web as “a set of sites that have more links (in either direction) to

members of the community than to non-members” (2000, 150). In the same spirit Robert Ackland and Rachel Gibson perceive web communities as structures of websites such, that if someone enters one of the member pages and follows the outgoing links, he or she is more likely to visit other members of this community rather than others (2004, 2).

Another point of view is embedded in web mining aiming at answering queries in search engines and finding structure of knowledge in the web. One of the basic assumptions is that all the information in the WWW may be divided into topics and sometimes it is assumed that such a collection of web pages sharing the same topic is a web community (e.g. Toyoda, Kitsuregawa 2001, 1). Jon Kleinberg devised one of the most known and most often applied structures, which supposes that every topic consists of authoritative pages, possessing most definitive knowledge and of hubs, i.e. pages which most often link to authorities (1999). Similarly, Ravi Kumar and his associates stated that knowledge bases (carriers of topics, i.e. sites on the same topic) are structured as bipartite cores, i.e. consisting of two sets of nodes, in the which all the nodes from one set link to the other (Kumar et al 1999, 640).

These examples show that communities are perceived either through the perspective of structure of links or through semantic cohesiveness. The second approach towards a community is deceptive from the sociologist point of view because it is not so that all the people who interact on the same topic constitute a community. It may be quite the contrary; people communicating on a given topic may not share one point of view, what may result in the emergence of antagonistic communities, which apart from the same topic do not share other features with each other.

## 2. Discourse and interaction on the web

Sharing the same topic by a set of web pages in a given language is actually embedded in discourse – a particular usage of language. It is usually characterized in three dimensions: 1)

interactions, i.e. events, in which agents engage in a verbal exchange, 2) linguistic content of that exchange, i.e. ordered string of words with their associated syntactic and prosodic structures and 3) the structure of information that is presupposed and/or conveyed by the interlocutors during the course of the discourse event in view of the specific linguistic content of the exchange (Lewis 1979, Grosz and Sidner 1986, van Dijk 1985).

How to measure interactions in the WWW? If we constrain ourselves to perceive interaction as an action which occurs when two or more objects have an effect upon one another, then links may serve as the indicators of an interaction. In most of works on the topology of links, they are usually seen as indicators of (citation) importance (Brin and Page 1998) or authority (Kleinberg 1999), but this approach constrain possible interpretations. Why not put a link on my page to point my enemies, for example? Excluding the case of web robots, it is usually so, that if I put a link to site A to site B, it usually means that I have seen that content of that page B and for some (any) reason I express this fact on page A. I may say that the content on page A resulted in a link on page B and a link on page B effects by expressing a higher visibility of page A, which actually is an interaction.

A set of web sites possessing a similar set of words constitute a topic in the Internet discourse. If among these sites some are densely linked with each other, they create a topic based community. The structure of these communities may be grasped by the network analysis of their interconnections, what is especially interesting when comparing several discourses. Different discourses may be featured by different organization of links. Below there is presented the abstraction and link analysis of the websites on three different discourses: politics, Christianity and culture in the Polish Internet.

### 3. Crawling the web

The aim of the crawling was to recreate three numerical networks of the domains which would be most important among their discourses. Nodes represent Internet domains, i.e. collections of web pages residing under individual URLs. Every domain consists of at most 200 single web pages inside a given domain. The vertices represent the sums of links between the domains.

In order to avoid the problem of a topic drift, i.e. to get only appropriate domains (Ackland, Gibson 2004, 6), it was necessary to create a topic based crawler, appropriate seeds (i.e. initial websites) and sets of keywords to evaluate crawled pages. The algorithm was written in Perl with help of MySQL database. Its mechanism was as follows. There was a dynamic list of domains (actually a table in SQL), where initially the seed was put. The crawler entered the following domains and apart from storing the links, it checked the occurrence of the keywords and the Polish fonts (in order to drop out foreign sites). It crawled 200 inner links in the order of occurrence. If on the first three web pages there appeared Polish fonts and at least three occurrences of the words from the key word set, it followed on. Every crawled domain had attached the number of all the occurrences of the key words which acted as its weigh. This weigh was attached to every outgoing link coming out of a given domain. Having crawled all the initial domains the crawler followed the link which in a given cycle of a loop had attached the highest weigh.

There were three different seeds for three different discourses. The choice for politics was the set of 26 websites of political parties as appeared on Polish Wikipedia in October 2006. For Christianity the list consisted of possibly most diverse orientations represented on the sites and had 17 web pages. For the culture I picked subjectively 7 most important cultural portals in the Polish web.

As for the keywords words I decided to apply various strategies. In case of politics I have taken the 98 entries from a political thesaurus (Chmaj, Sokół 1999). Because Polish language is inflectional, I cut off the suffixes of the nouns and took into consideration all possible forms of their lexemes. Eventually, it turned to be a bad choice. In Polish language word building is quite sophisticated, so many words occurred out of these lexemes, which had nothing to do with the preferential meaning (e.g. Polish *praw\** – root for *law* occurred among other as *oprawca* – a torturer). In case of religion the initial seed was crawled in order to count word frequencies. First, in the set there were excluded grammatical words, like pronouns or prepositions. All the words which occurred more than 400 times were taken into consideration. There were 214 of them. The seed for culture was examined in view of categories acting as the catalogues of their content. All told, there were 316 combinations of 46 different words for culture.

The set of the keywords is essential in the topic distillation. If a researcher chooses the words in advance, i.e. she decides which words constitute the topic, the outcome may be surprising. When I conducted similar research basing on different sets of keywords I realized how much it depends on the words. In one of the early research I inserted into the keywords Polish version of the word *freedom*. In consequence the crawler got fixed on homosexual sites, what was the proof that in the Polish Internet it is homosexual movement which speaks on freedom most intensely. As for the discussed set of keywords for politics derived from a political thesaurus, it does not seem to be the best solution, for there were many sites of public administration and local governments which were excluded in this case as the politics was not discussed on them. Also cutting off the suffixes was a bad choice, and in consequence there were only 67% of good sites. It seems that the best choice was in case of Christianity where the outcome of crawling consisted of 87% of good choices, i.e. the sites where Christianity was discussed. The rest of the domains consisted of foreign sites possessing Polish subpages.

In case of culture the evaluation is harder, because of many sites which are not directly devoted to culture, but discuss their matters. Culture is the vaguest discourse among the rest and it is hard to state, whether something is or is not culture. All in all, there were only 8% of the domains which apparently dealt with other topics and had no references to widely understood culture.

In order to be able to describe the derived communities I have categorized the web sites according to their form. Initially the categories were as broad as possible so as not to lose any slight differences between them. Eventually there remained the following categories:

1. A multi topic portal – a big site on various topics, allowing different communication services, such as blogs, forums, newsletters, and so on, possessing an internal search engine
2. A single topic portal – the same kind as a web portal, but focusing on one kind of a topic
3. A news site – a site focusing on one kind of topic, not offering any communicational services, with the domination of dynamic content and an internal search engine
4. A static page – a site offering (mostly) static information on an institution, topic, hobby, etc, but not on persons.
5. A blog – a site with dynamic content being sorted from the newest to the oldest
6. A personal website with a blog
7. A personal website without a blog
8. A forum
9. An E-shop
10. A link directory

#### 4. Describing a discourse

Among social scientists discourse is often perceived as a semantically cohesive structure, influencing social actors in some particular way. From a communicational point of view the communication scheme of a discourse would seem as follows: actors create semantic cohesiveness through interactions, and then the discourse influences the actors. But it so, that different actors create different points of view on discourse on the one hand, and on the other, discourse may influence actors in different ways. That is why it may be interesting to look at the structure of actors. Network analysis allows checking the cohesion of actors' interconnections.

Specific measures can give some insight and are well described by Peter Monge and Noshir Contractor (2003). Cohesion may be thought of in terms of solidarity and self-consciousness. These are well described by discourse reciprocity, transitivity and density. Monge and Contractor compare them to sociological theories: reciprocity endorses exchange theory, transitivity confirms balance theory and density affirms collective action theory (2003, 56). However, in case of the hyperlinks inside a discourse, it may be a bit different. Density is a very appropriate indicator of the level of self-consciousness. The denser the discourse, the more people know each other. Inside given discourse, if there are many links between topic communities, the more self-conscious is the discourse itself. If there are hardly any links between the communities, the discourse is antagonistic. Other information – the number of links coming to and from given domains accounts for reciprocity. If a given cluster is highly reciprocal, there may be assumed higher uniformity among the authors of links. If there is a case of a dense cluster with small reciprocity, on the other hand, there may be faced pluralism. Transitivity is another interesting measure on self-consciousness stating on the grounds on the link topology, that if there are three domains A, B and C and A links to B and B links to C,

whether there is link from C to A. If there are many such A, B, C triads, then transitivity can tell a lot about self-consciousness, but in case of few such triads, it is not very sufficient measure. One other single measure describing network authority is the betweenness centralization index. It measures the distribution of most central nodes in view of their betweenness and can tell, whether the stratification of the domains is horizontal or vertical.

More information can be added by looking at the vertical and horizontal modes of communication and distribution of forms of web sites in given communities. The link distribution, first of all, can tell us, whether there is vertical mode of communication (few authorities and others passively listening) or horizontal mode of communication (authority is biased among all the domains). In general the link distribution follows the power law, which state that the probability  $P(k)$  of occurring  $k$  links equals  $k^{-\gamma}$ , where  $\gamma$  equals more or less 2,1 (Barabási, Albert 1999). If  $\gamma$  exponent is high, the link distribution is biased. If it is small, then link distribution is very uneven. Authority distribution in a given discourse can also be measured by the variety (or its lack) of the possibilities of clustering a discourse. If there are few possibilities of dividing a discourse into communities, it is much more rigid. The best device for this evaluation is a community search algorithm. In general it tries to divide a network into given number of clusters, i.e. the collection of nodes which are heavily interconnected inside and have no connections to other clusters. These algorithms usually provide the quality of such divisions – the difference between the ideal model and empirical data.

Initially I applied the faction analysis and tried to interpret the distribution of quality values, but it soon turned out, that in case of analyzed clusters the quality distribution starts with the high number for small number of clusters and logarithmically diminishes with the increase of the outcome clusters. Then I tried the algorithm devised by Michelle Girvan and Mark Newman (which I will refer to as NG). Its mechanism in general is as follows: given the

number of desired clusters, it looks for the relations featuring the biggest betweenness (connecting nodes which are not connected by other relations) and deletes them, checking if it sufficiently divides a network into the given number of unconnected parts. Basically the value of quality is checking the difference between the numbers of empirical connections inside and between clusters with the ideal type where there is the smallest possible number of connections between the clusters and the biggest inside clusters (Girvan, Newman 2004).

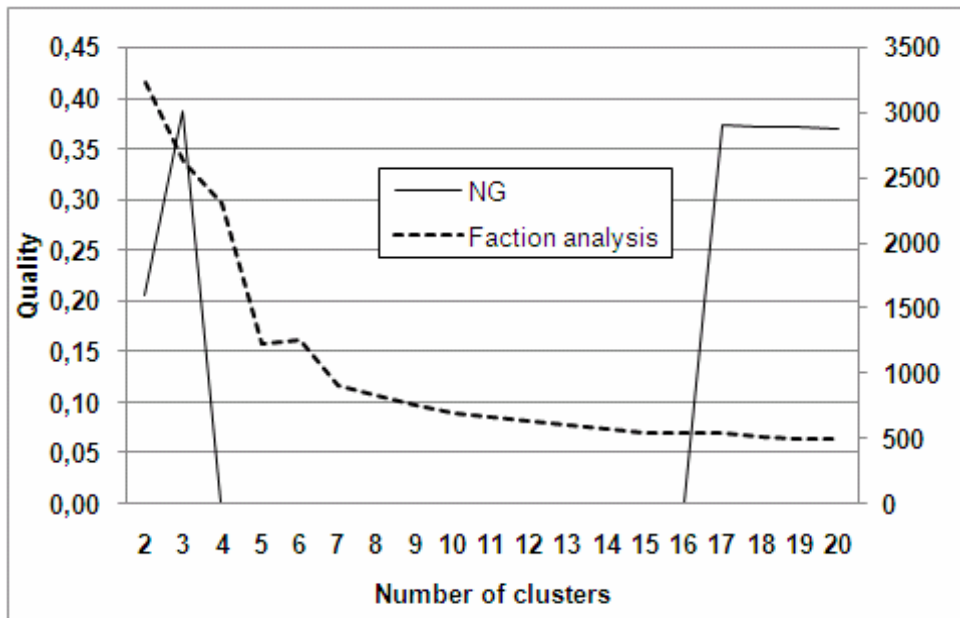
Another useful information is the proportion and interrelation of forms of websites in given discourses. In general there are two main conclusions which may be drawn on it. First of all, the domination of single topic portals, news sites and static pages on hobbies and institutions can tell that its users are oriented towards collectivism, i.e. communication is oriented towards institutions, topics and so on, while the domination of blogs and personal pages indicates individualism, where more stress is put to individuals. Second, if most of the domains allow for various communication tools, posting comments and send e-mails, then the discourse is much more interactive and thus less authoritative and the dominating model of communication is multi-directional. On the other hand, if most of the domains are static and disable the possibility of interaction, the communication is much more passive and thus site guests remain in the roles of recipients of the offered content in the given discourse.

## 5.1 Politics

As for politics, the outcome was not obvious. In older democracies, the political system is biased towards a two party system, which may result in the structure of two main clusters as was in case of the research done by Natalie Glance and Lada Adamic (2005). In Poland, however, the situation is much more complicated and the common division into left and right wings serves as the ideology brand mark rather than the analytical tool, explaining any

regularities. The distribution of quality results for Faction Analysis and NG algorithms are presented in the figure 1.

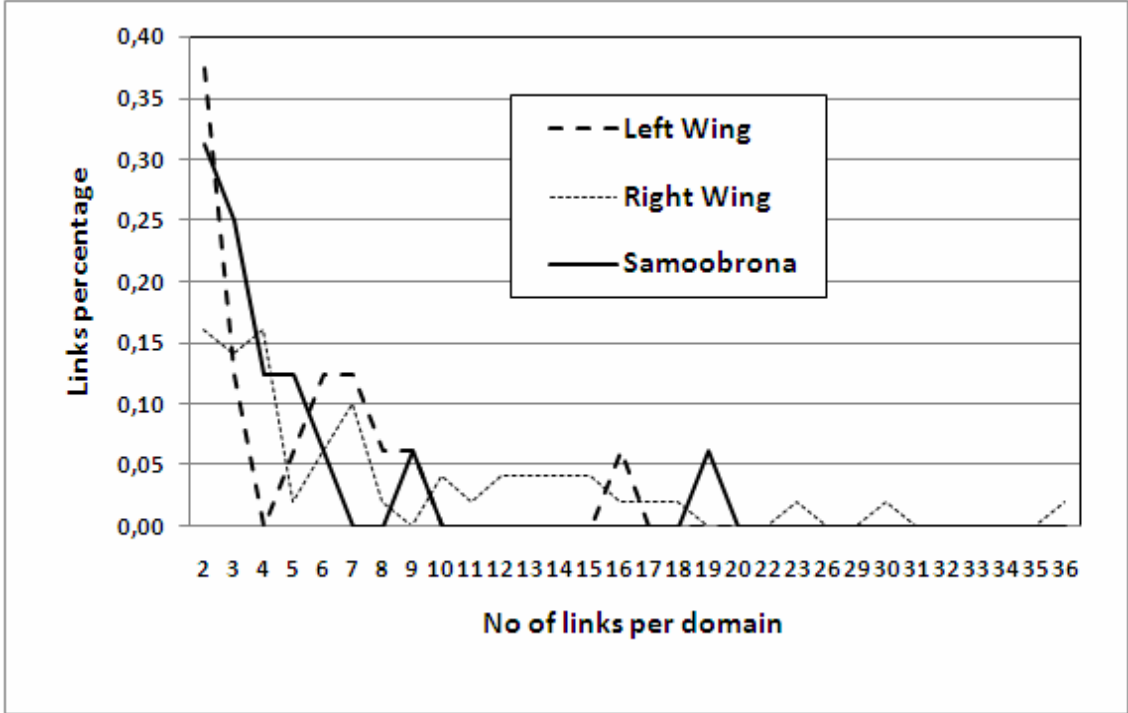
**Figure 1. Quality of cluster division by NG algorithm for politics.**



Taking a closer look at the results shows that the biggest probability of cluster divisions allows for two or three clusters, whereas bigger number of clusters is not predicted by the NG algorithm (apart from big number of clusters which are not interesting from a general interpretation). Looking at the three clusters possibility ( $Q = 0,387$ ) reveals that the political network is divided purely on the ideological basis into the right wing (64 domains), left wing (24 domains) and a graph of a populist party – Samoobrona (eng. *Self defence*) (20 domains). The names are derived from the names of the domains with the biggest number of links in the clusters. It is also worth mentioning that there are only 26 links between the clusters. Such a division reflects political situation in Poland in the end of 2006. It is endorsed by the fact that Samoobrona is related to the right wing – these clusters are more connected

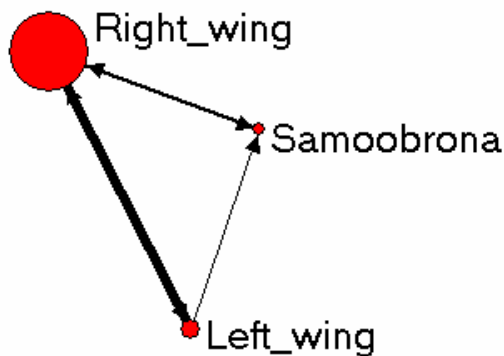
and are united in case of the NG division into 2 clusters and being opposed to the left wing – Samoobrona has been in coalition with the rightist government since then.

Figure 2. Link distribution in the political network



Rigid possibilities of clustering show that political discourse is heavily and clearly divided. What also accounts for this observation are both – a very small number of 26 links between the clusters and a relatively small exponent  $\gamma$  (1,6) for the whole politics, 1 for the Right Wing and 1,5 for Samoobrona, what shows that these discourses are structured in a vertical way. It is supported by the fact that when counted the cluster betweenness (i.e. the number of links leading to clusters which wouldn't be connected otherwise), the only value is attached to the Right Wing, as it has minimal references to Samoobrona.

Figure 3. Political clusters



The derived clusters have different structures when authority is taken into account. The most central network is of Samoobrona (Betweenness centralization index: 73,42%), even though its  $\gamma$  exponent equals 1,5. The only key players are: the central site of the party which is in fact a news site and a static page of its leader. This cluster also features comparatively high reciprocity, but it is only a proof of the fact, that central domains are aware of its tendrils and the tendrils refer to the central domains. Still, the tendrils are hardly aware of each other. Self-consciousness of the actors in this network should be higher than in case of other political options, as the link transitivity equals 1,14% in comparison to 0,21% of the Right Wing and 1,14% of the Left wing; this relation, however, tells nothing because of the small number of triads in this network. The high level of authority is also reflected by the vertical mode of communication; Samoobrona consists mostly of the personal pages without blogs and static pages of sub organizations. Static communication lessens the possibility of interaction and in consequence makes it more vertical.

The most typical political web graph maintaining its diversity and different modes of communication is the Right Wing. Although it has the smallest  $\gamma$  exponent, which equals 1 (what suppose very unequal distribution of links), its centrality is of moderate degree (Betweenness centralization index - 35,53%). It has the highest number of blogs (88,9% of all the political blogs) and is focused around the most prominent websites: two sites of political

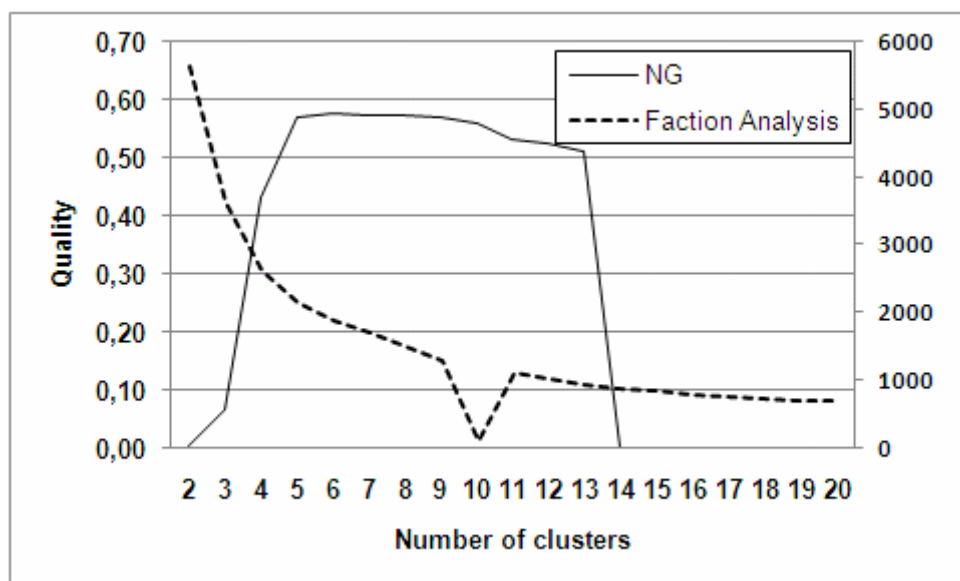
parties (Law and Justice and The Union of Real Politic), a political portal (prawica.net) and the most prominent rightist blogger, Galba (who has the highest number of outgoing links). The difference between these sites is interesting; while the two sites of political parties have a big number of incoming links, they hardly link to anyone else. The case of prawica.net and Galba is utterly different. This cluster is actually hardly cohesive – although it creates one cluster, the values of both reciprocity and transitivity are relatively small, what accounts for pluralism - the fact, that the rightist politicians are divided and not open towards each other. Particular rightist options are represented by portals, news sites forums and blogs. Both the liberals and conservatives inside the rightist cluster possess all forms of websites.

The left wing is most surprising for the fact that comparatively to the right wing it is actually smaller than in other European countries (Ackland, Gibson 2005). It may be said that it's the least centralized political graph (Betweenness centralization index 30,22%) with the most even link distribution ( $\gamma = 2,1$ ). The whole network is centred on the main portal (lewica.pl) and the number of smaller news sites and smaller single topic portals, while there are hardly any blogs. It is interesting, whether this fact reflects the collectivism of leftist ideology. If that would be so, the prevailing number of blogs in case of the rightist ideology would suggest its individuality. Indeed, the leftist network is more transitive and denser than the rightist one. It should also be stated that leftist portals are much more essential than rightist, have much more references and contain more information than the rightist Internet in Poland.

## 5.2 Christianity

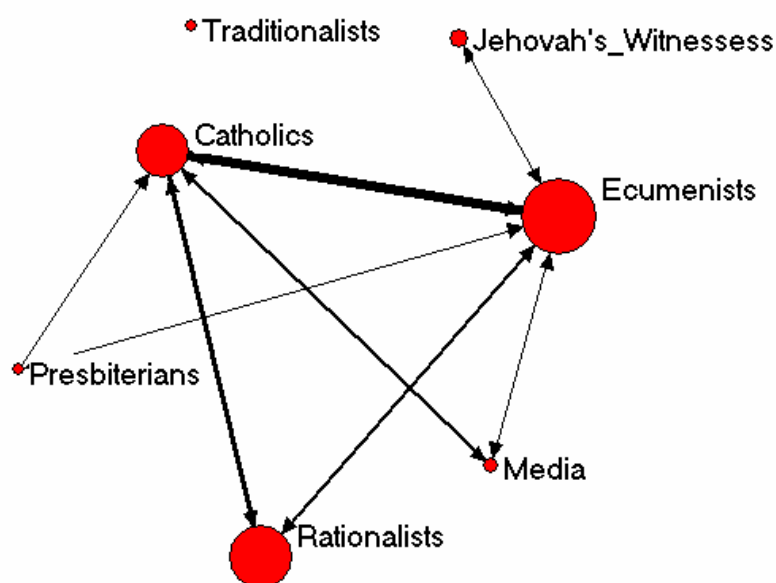
The Christian network seems to be the inversion of politics when looking at the possibilities of clustering. The seed was supposed to include all the possible variations of Christianity in Poland, starting with the institution of a Catholic Church, coming through Protestants and finishing with the Greek Church in Poland. The biggest quality value for clustering with NG algorithm is for 7 clusters (0,577) and divides the network into: Ecumenism together with the protestants, the liberal wing of the Catholic Church and the Greek Church (40 domains), strong core of the Catholic church (27 domains), rationalist anti-church movement (33 domains), Christian media (4 domains), Jehovah's Witnesses (6 domains), Presbyterians (2 domains) and Catholic traditionalists basing on the teachings of archbishop Marcel Lefebvre (2 domains).

Figure 4. Quality of cluster division by NG algorithm for Christianity.



The biggest Quality of NG algorithm regards 7 clusters and the following possibilities focus on the diverse subdivisions between Protestants and liberals in the Polish Christianity with the slight differences in the division of Catholic media. The ambiguity in this division may be due to weaker tradition of the protestant wing of the Christianity. Still, it has stronger traditions than in case of politics, which has stricter divisions. Probably, the value of quality is in itself insufficient information on the possible divisions in a network. What counts as well is the number of similar possibilities, showing the variation of communicational ties and the fact that there are 101 links between the clusters.

**Figure 5. Christian network**



As far as the level of clustering is considered, Christianity resembles politics in a way. The quality results of the NG algorithms are even more precise (0,577 for Christianity while only 0,387 for politics), yet there are some differences. First of all, the relation of links between clusters and the number of clusters is much bigger in this case (Christianity has 7 clusters and 101 links between them, while Politics consists of 7 clusters and 26 links joining them), and

second, what is obvious when looking at the overall network, the Ecumenist web sites are connected to the whole network and thus enable the dialog between almost all the varieties of Christianity and maintain the dialogue. This entire network is also apparently less authoritative than Politics. Its overall betweenness centralization index equals 12,69%, while the same value for Ecumenism is 24,98%. That is why maybe it is useful to call such a structure a dialogical discourse.

The most important nodes of the general Christian network are the Ecumenists, Catholics and rationalists. Although Poland is supposed to be a Catholic country (at least on the level of declarations), in the Polish Internet Ecumenism is the most prominent Christian ideology (i.e. liberal Catholics and Protestants). It is also Ecumenists who maintain the dialogue between all the varieties of Christianity. Their network is centred on three main single topic portals: [ekumenizm.pl](http://ekumenizm.pl), [luteranie.pl](http://luteranie.pl) and [kosciol.pl](http://kosciol.pl), while the majority of domains are static pages and news sites of foundations, institutions and small branches of Protestantism. Small density accounts for its pluralism and probably not too big self-consciousness. Still, there is a considerate number of reciprocal ties (0,16), what makes this cluster quite interactive, at least in comparison to politics.

Catholics are similar to Ecumenists in a way. Its  $\gamma$  exponent (1,4 for Ecumenism and 1,5 for Catholics) and the level of betweenness are similar (both about 25%) and as such both of these partitions are similar in its authority. However, on the one hand it is Ecumenism which maintains contact with the rest of Christian domains, and on the second, while Ecumenist sites are full of essential information on the faith, it is Catholics who possess more portals and enable the contact of individual Christians in the Internet. There are also more online shops in this case. Thus there are more forums, mail boxes and other informational services. Still in both of these clusters there are few blogs, and if they do exist, they are devoted to the matter of faith. Religion features collectivism instead of individualism. It also

endorsed by the fact that all the domains are first of all linked to the collective portals focused on faith – key players in the Christian discourse.

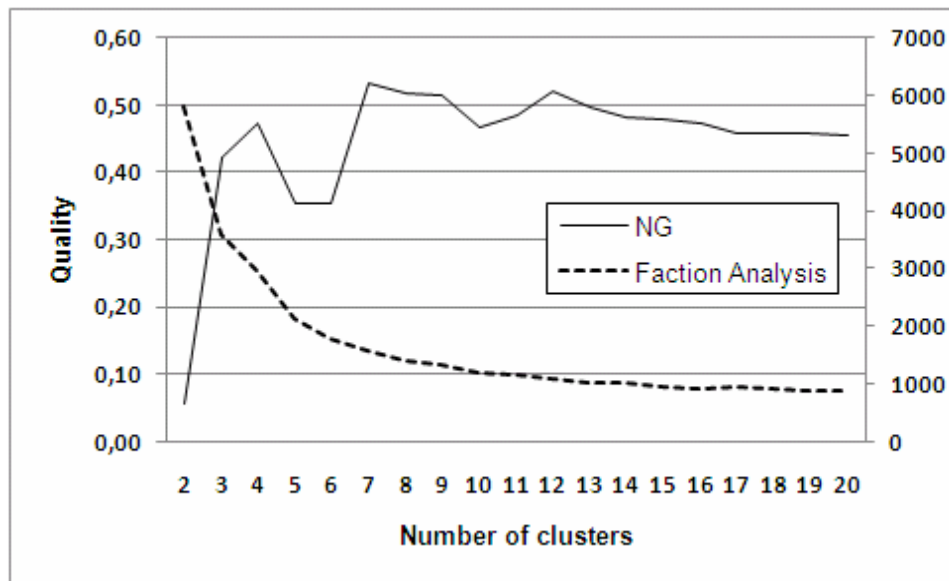
The most surprising cluster is the one representing the Polish Rationalist Society, which discuss the matters of faith, yet in a very opposing way. It is the most central discourse, next to Samoobrona, although the Betweenness Centralization Index does not tell it at all. All the domains are only referred to by 4 main sub domains of the portal racjonalista.pl and they do not link back. Yet the portal is one of the biggest and most well known in the Polish Internet, full of communication possibilities and web information. It is also very professional, as far as philosophical discussions are concerned. It is hard to answer the question, whether this discourse is authoritative, because of the fact, that most of the rationalist discourse takes place in the mentioned portals, while the rest of the domains show only its references.

The remaining domains are the representatives of Christian media, Jehovah's Witnesses and Christian Traditionalists. They will not be described due to their small number, but it is worth mentioning that the domains of the last option are not connected to the Christian network at all. It seems that in the Polish Internet (at least on the level of this research) Traditionalists do not associate themselves with the dominant Christian institutions at all. Even Jehovah's Witnesses are referred to by the portal Ecumenism.pl.

### 5.3 Culture

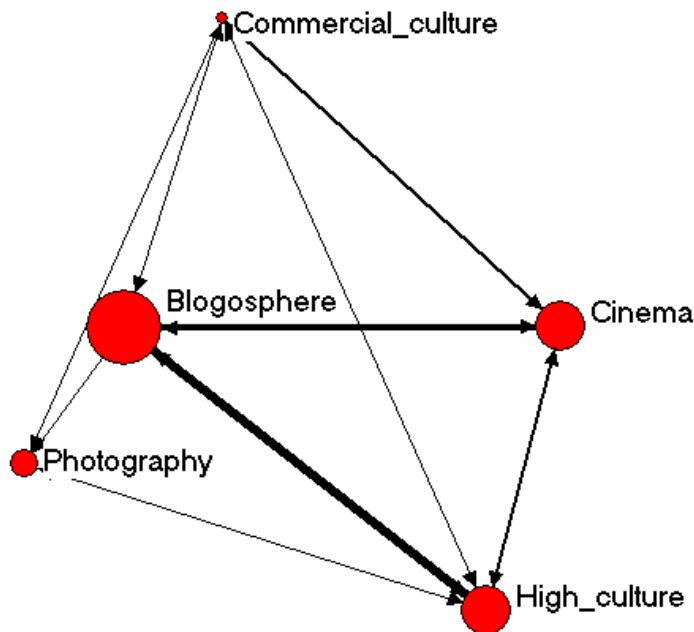
The last network – culture is even more different and hardly can be called a discourse in the same way as politics and Christianity. Unlike two previous networks, the distribution of quality is most ambiguous and gives few possible hints for interpretations. This discourse is most dense of all and, unlike in previous cases, the distribution of links is not always due to the content of the domains, but also to their form and consequently – communication modes.

**Figure 6. Quality of cluster division by NG algorithm for culture.**



The biggest quality (0,532) features 7 clusters inside cultural network (while 1 cluster features 3 unconnected sites and will be omitted in the description). It also represents the content differences in a most obvious way, but differs from politics and religion. This division is certainly not based on either ideology or kinds of arts. The first cluster is a very dense blogosphere understood in most direct way – collection of 33 blogs which are primarily oriented towards each other, rarely referring to the websites of artists or galleries. Another cluster (31 domains) consists of commercial movie production and has almost no connection with the rest of the cultural domains (it was pointed by one blogger and e-shops). There are also 24 nodes representing the most prominent art in Poland, containing websites of famous galleries and “high brow” intellectual portals. Another 16 nodes reflect the connection between commercial and artistic photography and seem to be more oriented towards themselves rather than the rest of the overall art in the Internet. 11 domains reflect some institutions, blogs and portals which connect the world of commercial and independent art. There are also remaining two clusters consisting of three domains, loosely interconnected with the rest of the network. Strikingly enough, the clusters are interconnected by 957 links.

Figure 7. Cultural network



What is most interesting is the fact, that the distribution of cultural links does not exactly follow the power law. The maximum number of links per domain is 14, but this network is so dense that the percentage of links per domain does not diminish with the increase of its number per domain, especially in case of a blogosphere and high culture.

The cluster which I call a blogosphere consists of 23 blogs, 5 personal pages without a blog, four news sites and a static page. Although its betweenness centrality states that it should be more authoritative than the other described clusters (37%), this information is misleading due to the linear distribution of the percentage of number of links per domain. Relatively it is one of the densest (on the level of 0,17) and most reciprocal (0,24) clusters. If that is so, its pluralism should be relatively small, when people first of all refer to each other and the level of self consciousness is high. However, a high number of blogs suggest individualism, but this individualism is oriented towards other individualist. Blogosphere is connected to other clusters (first of all cinema and high culture), as its cluster betweenness equals 3,8%. Interestingly enough, the high level of density is primarily maintained by the feminist movement, apparent in the Polish cultural discourse in the Internet.

High culture is the cluster representing first of all Polish art, galleries as well as main portals and news sites dealing primarily with painting, sculpture and modern art. Its core is the main single topic portal [obieg.pl](http://obieg.pl), possessing most of the in and outgoing links, together with a few other small portals and a few sites of art galleries, among others [csw.art.pl](http://csw.art.pl) and [raster.art.pl](http://raster.art.pl). The majority of the domains are static pages or news sites. When looking at this network, it is important to remember that actually this discourse is rarely created by artists themselves, but by galleries, journalists or publishing houses. That is why this is not a discourse of artists but a discourse on art. Thus, apart from the differences in the forms of prevailing websites, its structure is similar to the one of a blogosphere – it is centred on the most prominent art portals and is similarly authoritative. Still, it is a little bit less dense and much less reciprocal. It is possible that the discourse on art means less self-consciousness and smaller uniformity, implying that the art galleries are connected mostly by portals and news sites.

The next cluster consists of the sites regarding movie production; however most of them are sub domains of the biggest Polish movie portal, [filmweb.pl](http://filmweb.pl). Still, it is connected with all other clusters apart from photography. Because this structure is not emergent, but created by the authors of [filmweb.pl](http://filmweb.pl), it will not be described herein. However, it is interesting, that apart from the strict movies sites, there is a connection to other clusters by two main Polish magazines, *Polityka* and *Wprost*.

Photography is the least connected cluster with the overall cultural network. It consists of static pages, news sites which are connected in a very sparse way and in general does not allow for much interactivity between its users. Due to the small number of blogs it is possible that it is not too individual (there are hardly any pages of individual photographers) and it's even less reciprocal than high culture. On the other hand it is hardly centralized what accounts for the fact that it is not a cohesive and self conscious discourse. Of course it's hard to



**Table 2. Distribution of forms of websites in the discourses.**

	Multi topic portals	Single topic portals	News sites	Static pages	Blogs	Personal pages with blogs	Personal pages without blogs	Forums	E-shops	Directories
<b>Politics</b>	3	20	15	29	18	4	16	6	0	0
Right Wing	2	11	9	16	16	1	3	6	0	0
Left Wing	1	9	4	6	2	0	2	0	0	0
Samobrona	0	0	1	5	0	3	11	0	0	0
3 excluded	0	0	1	2	0	0	0	0	0	0
<b>Christianity</b>	4	29	22	40	5	1	5	0	7	1
Ecumenists	0	7	10	19	2	1	0	0	0	1
Catholics	0	12	2	8	1	0	0	0	4	0
Rationalists	4	5	6	8	2	0	5	0	3	0
Presbyterians	0	0	1	1	0	0	0	0	0	0
Media	0	2	0	2	0	0	0	0	0	0
Jehovah's Witnessess	0	1	3	2	0	0	0	0	0	0
Traditionalists	0	2	0	0	0	0	0	0	0	0
<b>Culture</b>	1	9	19	47	26	5	6	1	4	0
Blogosphere	0	0	4	1	20	3	5	0	0	0
High culture	0	5	6	11	1	1	0	0	0	0
Commercial culture	0	2	1	4	3	0	0	1	0	0
Movies	1	0	4	22	1	1	0	0	2	0
Photography	0	2	4	8	0	0	0	0	2	0
3 excluded	0	0	0	1	1	0	1	0	0	0

## 6. Conclusions

Mining the web in order to describe and analyze particular discourses may provide some insight into the communicational structures. Probably it may be used to draw conclusions on a society on the grounds of the topology of hyperlinks, under the condition that it is remembered that, first, the discourse is created by people using the Internet and thus not the whole society, and second, that such discourse is created by people who talk on it, not necessarily by the key actors in a given discourse offline. That is why such a distribution of links may describe discursive orientations in the society of Internet users. However, the more common the Internet is in a society, the more it reflects common attitudes in such a society.

Sole hyperlink topology reflects patterns emerging from interactions, not yet allowing for sophisticated content analysis. But such patterns of interactions allow grasping the communication structures of a given discourse. Drawing on the example of presented

discourses there may be presented a few major ideal types of their structures, which I call antagonistic, dialogic and individualistic.

Antagonistic discourse consists of a set of clearly divided ideological clusters, which rarely link to each other. Ideally, every cluster should be very dense (allowing for self consciousness), reciprocal (so as to feature uniformity) and triadic (to be more balanced). It also should be collective, i.e. people should communicate content on behalf of groups or institutions or ideas which they endorse. People embedded in such discourses aim at pushing forward their ideas without arriving at consensus.

Dialogic discourse is similar to antagonistic but trying to maintain the dialogue between all the diverse ideological fractions within. There should be some place where every one can exchange ideas with followers of other orientations. In this case there are many links connecting highly divided clusters, allowing for the interaction with different people. Here constant dialogue is a value.

In both the antagonistic and dialogical discourses all the communication is centred on main single topic portals, allowing for communication on a given topic. They are encircled by static pages of the advocates of given ideological options. If there are individuals, they are always collectively devoted to the case (at least their visibility is maintained by their devotion to the case). Every ideological option is structured in a manner reflecting its interpersonal relations.

Individualistic discourse is not focused on ideas, but on people who are attracted to each other by means of the will to communicate. Clusters in here are not so important because what counts is the domination of links between the clusters. When people communicate, they do it on their own behalf; hence there are many blogs and personal sites. Because of the lack of highly visible clusters, the betweenness is very high and due to the pluralism any many points

of view, the reciprocity remains rather low. Self consciousness of such a discourse is not that obvious.

These models remain as such but may be a good introduction to the research on the nature of ideas and trends online. Due to such an approach it is possible to say, whether a given phenomenon is well established in the social consciousness and what the attitudes of people connected with it are.

## References:

Ackland, R., 2005, Estimating the size of Political Web Graphs, revised version of paper presented to ISA Research Committee on Logic and Methodology (RC33), 17-20 May 2004, Amsterdam, on: [http://acsr.anu.edu.au/staff/ackland/papers/political\\_web\\_graphs.pdf](http://acsr.anu.edu.au/staff/ackland/papers/political_web_graphs.pdf).

Ackland R., Gibson R., 2004, Mapping Political Party Networks on the WWW, a paper presented on *Australian Electronic Governance Conference*. Melbourne, University of Melbourne, on: [http://voson.anu.edu.au/papers/political\\_networks.pdf](http://voson.anu.edu.au/papers/political_networks.pdf).

Adamic, L., Glance N., 2005, The Political Blogosphere and the 2004 U.S. Election: Divided They Blog, on:

<http://www.blogpulse.com/papers/2005/AdamicGlanceBlogWWW.pdf>.

Barabási, A. L., Albert R, 1999, Emergence of scaling in random networks, *Science* 286, 509-512.

Brin S., Page L., 1998, The Anatomy of a Large-Scale Hypertextual Web Search Engine, on <http://infolab.stanford.edu/~backrub/google.html>.

Chmaj, M., Sokół W., 1999. *Polityka-Ustrój-Idee; leksykon politologiczny*. Lublin: Morpol.

van Dijk T., 1985, *Discourse and communication: new approaches to the analysis of mass media discourse and communication*, de Gruyter, Berlin, New York.

Flake G., Lawrence S., Giles L., 2000, Efficient Identification of Web Communities, paper presented on Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Boston, MA.

Girvan M., Newman M., 2004, Finding and evaluating community structure in networks, *Physical Review E* 69 026113, on <http://arxiv.org/abs/cond-mat/0308217>.

Grosz B., Sidner C., 1986, Attention, Intensions and the structure of discourse, *Computational Linguistics* 12, 175-204.

Kleinberg J., 1999, Authoritative Sources in a Hyperlinked Environment, *Journal of the*

ACM 46(5), 604-632.

Kumar R., Raghavan P., Rajagopalan S., Tomkins A., 1999, Extracting Large-Scale Knowledge Bases from the web, Proceedings of the 25th VLDB Conference.

Lewis D., 1979, Score keeping in language game, in: Bauerle R., Egli U., Von Stechow A., Semantics from different point of view, Springer, Berlin.

Merton R., 1957, Social theory and social structure, Free Press, Glencoe, Ill.

Monge P., Contractor N., 2003, Theories of Communicational Networks, Oxford University Press, New York.

Pennock D., Flake G., Lawrence S., Glover E., Giles L., 2002, Winners don't take all: Characterizing the competition for links on the web, Proceedings of the National Academy of Sciences 99(8), 5207-5211.

Toyoda M., Kitsuregawa M., 2001, Creating a Web Community Chart for Navigating Related Communities, Proceedings of the 12th ACM Conference on Hypertext and Hypermedia, on [www.tkl.iis.u-tokyo.ac.jp/~toyoda/klieg/hypertext2001-toyoda.pdf](http://www.tkl.iis.u-tokyo.ac.jp/~toyoda/klieg/hypertext2001-toyoda.pdf).

Tönnies F., 1988, Wspólnota i stowarzyszenie: rozprawa o komunizmie i socjalizmie jako empirycznych formach kultury, PWN, Warszawa.