Empirical Investigation of the Blog space: the Case of European Pharmaceuticals

Emanuela Todeva  
University of Surrey, UK  
etodeva@surrey.ac.uk

Donka Keskinova  
Plovdiv University, Bulgaria  
keskin@evrocom.net

Abstract  
This paper provides the findings from a research project commissioned to assess the emerging body of information being published in web logs (blogs) in the field of European pharmaceuticals and European health care. The hypothesis behind this research was that the rise of blogs as a rich information source may provide a challenge to the traditionally held view of a ‘key opinion leader’ – could an anonymous blogger have as much influence over public debate as a recognised scientific expert? The results reveal that despite high volume of emerging blogs, only a small number are interlinked, and the emerging network configuration is a small core component and a large number of dyads, or short tails. The public opinion broadcasted in blogs shows differentiation related to specific health issues or pharmaceutical companies.

1. Introduction and conceptual background

The literature on on-line communities is still attempting to clarify the nature of virtual communities and the boundaries that draw distinctions between communities of practice, and communities as distributed intelligence systems. There is also an effort to draw a distinction between relational analysis and social network analysis (SNA) [12] and between social interactions, human-computer interactions and on-line communications [15]. Being a part of the internet, blogs by their nature allow interconnectivity, and there are expectations that they facilitate the formation of on-line communities [3]. The presence of social aggregation (i.e. interconnected social actors), ‘long enough’ connectivity, and a collection of on-line mediated personal relationships (i.e. interactions and direct one-to-one communication) are seen as evidence of these on-line communities [3].

As an internet technology, blogs are related to a group of interactive technologies developed for computer supported cooperative work in a network, collaborative editing systems and collaborative learning, or other collaborative on-line activities. In the design of these internet technologies the aim to facilitate collaboration is often substituted with the effect of the use of the technology – i.e. the collaboration itself. There is little attempt to discriminate – at what point of the use of an interactive technology, or broadcasting of information, and sharing of information with the public on-line becomes a form of collaboration. From substantive point of view, a similar question is when a group of communicating and interacting people become a community, and when group communication, group awareness and group coordination transform some group practices into community practices.

Blogs are seen both as a new publishing medium employed by and challenging the established mass media practices [7] and as a personalised expression and relational technique enabling individuals to speak for themselves [2]. In both cases, the expectations are that blogs can and do influence public opinion – either as an internet broadcasting tool, or as influence by key opinion leaders in communities with specialised interests. However, a crowd of people that do the same thing (i.e. broadcast and/or view online information) should be distinguished from coordinated action in a group of people with similar interests.

There are many questions that need to be addressed in this context - both at theoretical and empirical level. There are also many disciplines that are currently researching the internet space and contribute to our knowledge and understanding of the processes that take place and the impact of various enabling internet technologies. However, we have to distinguish clearly between technical connectivity and human behaviour in an interconnected environment (i.e. interconnected computers and web-pages vs. human interactions). It is common to substitute technological capabilities with the actual use of the technology. Although clearly blogs offer both access to content and connectivity, the presence of connectivity should not be used as evidence of a relationship, or other association in an affiliation network.

The typology of blogs, circulated in the literature [Lin] suggests that online personal diaries, collections of links to other sites, or on-line public forums devoted to specific topics – all remain on-line broadcasting tools. There is a need not only to confirm that the broadcasted information has been accessed by some audience, but also some form of interaction between a blogger as a sender and a blogger as a recipient of the information – commenting on it, or acting under
its influence. Simple single viewing of information is hardly any evidence of a relationship beyond simple awareness and association. Repetitive viewing or following the html suggestions from a blog however may be considered as affiliation – similar to the notion of group membership. Posting a comment can be considered as evidence of interaction and a ‘community’ relationship [3].

In terms of sources and effects of influence, we have to discriminate between a source of information (i.e. on-line publishing / broadcasting medium), potential recipient of information (all internet users that have online access to the source of information), and the interaction between the source and the recipient (evidence of impact of the information, or a reaction by the recipient in response specifically to viewing the source [12].

Communication and interaction in the internet space is still not an evidence of a community relationship, as it is not an evidence of sociability and sense of belonging [15]. In spite of the universal connectivity of the internet, there is a difference between an observer to a community and a member of that community. Community membership for an aggregation of individuals can be attributed by ‘sharing’ particular beliefs, or practices, and by interactions. In the context of blogs, posting comments can be acknowledged clearly as evidence of a community membership [3].

Posting a hyperlink to another blog is merely a unilateral referral to another source of information in a distributed intelligence system, and hardly an effective relationship between different blogging communities. Although individual blogs can be interpreted as a community of bloggers, hyperlinks that connect these blogs can not be treated as evidence of ‘a blog community’, let alone of an online community of bloggers that participate in discussions in individual blogs. Some evidence of repetitive interactions or impact on behaviour are necessary – to support a ‘community’ hypothesis for the blogosphere.

In a community of practice, interaction is assumed, and it is assumed that members of that community participate in the ‘community activities’. In a distributed intelligence system there is simply dissemination of information with or without feedback (i.e. with or without interaction). In this context, it is safer to hypothesise communication relationships between blogs and bloggers, rather than community relationships. Co-authorship and co-editing in distributed intelligence systems are clear evidence of communication relationships, as well as evidence of membership. However, the existing definitions of a community with reference to a group of individuals who share intent, belief, resources, preferences, or needs [3], does not allow us to draw clear boundaries – who is a member, and who is not a member of a particular community. The sense of belonging is essential [15], and only personal statements can be an evidence of a community membership. Link analysis without text analysis can reveal communication relationships that can substantiate community interactions, but should not be subsumed as such.

In our research project on blogs we applied an anthropological approach, or attempted to collect and analyse information on a selection of blogs that address issues from a specific semantic field (European pharmaceuticals and European healthcare). We attempted to investigate the ‘social’ connectivity between these blogs in any form – either by hyperlinks (i.e. distributed intelligence systems), or by participation in on-line discussions (i.e. communities of practice). In addition, we looked at semantic connectivity, or dominant semantic relationships based on a selection of key words that demarcate our semantic field. Our approach to blog analysis has been informed mainly by social anthropology, communication theory, semiotics and organisation theory.

2. Methodology and selection criteria

Blog analysis at present is known as a method for data mining, where the main question is to identify cascading behaviour and to find patterns, rules, clusters, or outliers in the WWW and to speculate on the ‘potential’ spread of influence across blogs linked by referral URLs [9]. New algorithms for page ranking are among the issues that have attracted scientific attention [7], [8], [14]. Searching core social structures and cyber-communities on the web has led to a number of mapping results identifying homogeneous groups of blogs by topic [7]. More comprehensive analysis of blog content and behaviour has been offered in the context of specific issues, such as the political discourse around the American elections in 2004 [1], or music blogs [3].

Our selection of the semantic field of European pharmaceuticals and European healthcare identifies a segment of the blogosphere which is an intersection of commercial and public interest and represents on-line media for dissemination and access to information (as well as a tool for self-expression and sharing of personal experience). As such, it is expected that the content of the blogs will reflect both commercial and private interests. Being a broadcasting media, it is expected also that the content will reflect current events and some deeper underlying individual predispositions to healthcare.

We have chosen to work with the full population of relevant blog-URLs over a fixed period of time (January-June 2008) and with a thematic selection of blogs (blogs that have made a reference to at least one of our key-words identified as representative of the semantic field of European pharmaceuticals and European healthcare.

Blog analysis has been associated with blog search and web-mining where the data comes in three main types: content (text, images, etc.), structure (hyperlinks), and usage (navigation,
queries, page ranking, etc.), implying different techniques such as text, graph or sequence mining [16]. We differentiated from these approaches by developing an alternative methodology for blog analysis that employed content and structure analysis simultaneously, in order to evaluate blog impact as emerging associations in a specified semantic field.

For the purpose of this project we developed a comprehensive methodology for blog search and blog analysis which comprises of seven main steps, including building a comprehensive database with the full population of related blogs, cleaning of the database, evaluation of reference links and semantic associations within blog URLs, and mapping of semantic and connectivity ties at micro (within blogs) and macro (across blogs) levels.

We have adopted an agency approach to the internet where actors are either firms discussed in blogs, or critical health care issues (expressed by key words), or the blog-URL pages themselves. Our application of network analysis to heterogeneous networks aims to reveal the underlying structure of associations between different types of actors that can be interpreted as part of an emergent communication structure in a distributed intelligence system.

We did not test any hypotheses of influence by individual opinion leaders, as our preliminary observations suggested very little comments to blogs, or lack of strong evidence of ’community interaction’ between bloggers in our selected semantic field. We attempted to infer influence by looking at the network position of individual actors in various one-mode and two-mode graphs, where individual actor position is an expression of the set of dyadic relationships of that actor (micro-level), the set of relationships within the neighbourhood (mezzo level), and the set of relationships in the entire selected population (global macro level) [3].

2.1. Development of the selection criteria

We conducted text analysis of the news on European healthcare and European pharmaceuticals broadcasted in official on-line media between January-June 2008, and identified the ’search-keywords’ reflecting key events and dominant issues during this period. We grouped them in 6 distinctive groups (health, drugs, diseases, industry, regulation, region). These groups of key words demarcated the boundaries of our semantic field within which we looked at semantic associations, or dyads of key-words with relatively high co-presence in the overall blog space.

2.2. Selecting a blog-search engine

From a range of blog search engines we selected Google blog (http://blogsearch.google.com). The main justification for this decision is that Google blog produced the minimum duplications of URL pages on initial observation, with the maximum of total URL pages identified in a filtered query.

2.3. Search string

We formulated search queries that combined positive and negative filters with Boolean operators such as AND and OR. The positive filter contained three components:

- the scope of the research (pharmaceutical / healthcare)
- geophysical relevance (Europe, UK / England, France, Germany, Spain)
- one of the selected key words (we constrained the sample to the key-words for the search with a name of a pharmaceutical company - Pfizer, Glaxo Smith Kline, Sanofi Aventis, Novartis, Hoffmann La Roche, Astra Zeneca, Johnson & Johnson, Merck & Co, Wyeth, Eli Lilly, Bayer, Lacer, Bristol Myers Squibb, Shire Pharmaceuticals, Chiron Corporation, Chugai, Takeda, Teva Pharmaceuticals, Ranbaxy). The use of a company name individualised our queries and enabled us to build a comprehensive database that have mentioned at least one name and has minimum semantic noise.

2.4. Building the databases

For the purpose of this analysis we explored different methods of sampling the blogosphere. The final database was generated as a representative population for all available blogs that contain at least one key-word from our search criteria. We downloaded full blog details of all blog-pages that appeared obtainable. The automated reduction between ’visible’ blogs and ‘obtainable’ was 85% (see table 1).

Table 1. Population size

<table>
<thead>
<tr>
<th>Total available URLs</th>
<th>Total obtained</th>
<th>- less duplicates &amp; ’shells’</th>
<th>Unique pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total URLs</td>
<td>11824</td>
<td>2995</td>
<td>990</td>
</tr>
</tbody>
</table>

2.5. Cleaning of database

After filtering the majority of duplications by the search-engine itself, we cleaned the database at multiple stages. For this purpose we used observation techniques and formal techniques based on proprietary software for URL searches. The cleaning of the database passed through the following stages: cleaning of duplicate URL-pages; cleaning of ’empty-URL-pages’ with size < 2 KB information; cleaning of ’shell URL-pages’ that contain dictionaries, job-announcements, lists of URLs without text; and URL classifications, or adverts (see Table 1.) According to this procedure
we built a database with the full population of blogs that corresponded to our selection criteria, containing 990 entries. Out of this population we identified 358 blogs (or 36%) as an interconnected core that contains shared blogs referring to more then one company (key word), and 633 blogs (or 64%) as periphery – i.e. blogs related to only one company (key word) (observed as pendants on Net 1.).

2.6. Internet count of key blog-indicators (primary analysis)

We calculated four additional indicators per URL: size of URL in KB; Cross-reference between URLs in DB (as internal links); Cross-reference to other blogs (number of external links); and Number of occurrences of individual key-words per URL page. Some of these indicators were used for additional filtering of the data, and the final numbers were recorded after the cleaning process was completed.

2.8. Data analysis and mapping (secondary analysis)

Mapping the selected population of blogs involves mapping of commercial and private broadcasting by bloggers that stir discussions and share related news for a period of 6 months. The bloggers (expressing opinion), the blogs themselves, and the shared content represent different types of actors that form a heterogeneous information system driven by public and private interests. Our network mapping aimed to reveal the emergent structural concentrations of different types of actors, and their network position inferring potential source of influence.

The analysis of these different types of agents required a tool that can deal with heterogeneous systems of actors, and we have selected the approach for heterogeneous relational analysis [3]. This method allows to analyse co-occurrence relationships in a large data set at a dyadic level, which is not available using multidimensional scaling, or other clustering techniques.

For the network analysis we interpreted the ‘relationships’ or links between the following actors in one-mode and two-mode graphs:
- pharmaceutical firms and URLs in database (Net 1, Net 2). The similarity measure in this graph represents co-occurrence of pharmaceutical firms in blog content of individual blogs;
- pharmaceutical firms and key-words grouped in three groups Health, Drugs, Disease (Net 3 – for Health). The similarity measure in this graph represents significant ties based on co-occurrence of key words and names of pharmaceutical companies in blog content of individual blogs;
- interconnected key words in specific semantic fields (i.e. semantic blocks) (Net 4 – for ‘Health’). The similarity measure in this graph represents co-occurrence of key words in blog content of individual blogs.
- relationships between URLs based on cross-reference between the pages (internal links between URLs in database) (Net 5). The similarity measure in this graph represents cross-references between blogs.

For Nets 1, 2, and 5 we have used the absolute value of ties, and for Nets 3, and 4 we have used value based on standardised residuals (Chi-square) [13].

3. Overview of results

The results from the blog analysis are grouped in 5 main sections: mapping the blog-space of European healthcare and pharmaceuticals; mapping the key actors in this space; mapping of relationships between blogs; mapping of the topics on which blog-participants publish (semantic analysis of emerging associations); mapping of the potential impact.

3.1. Mapping of the blog-space

The first two maps show distribution of URL-pages and their association with a particular pharmaceutical company. Net 1 shows that Pfizer and Novartis are quite central to the selected field, and Pfizer, Bayer and Novartis exhibit unique profiles, as there are no other companies located in a close proximity to them. However, calculated centrality measures indicate that the companies with the most similar referral profiles are Pfizer (DC [Degree Centrality]=611, GlaxoSmithKline (DC= 556), and Novartis (DC= 483). In addition, there are three sub-groups of companies that share a significant number of URL-pages. These are: Johnson & Johnson with Ranbaxy and Teva; Sanofi Aventis, Hoffmann La Roche, Merck & Co, and Bristol Myers Squibb; and GlaxoSmithKline with Takeda, Eli Lily, Shire, AstraZeneca, and Wyeth. The meaning of these sub-groups is that there is a significant number of blogs that compare and contrast these firms.

The strongest connections between firms and URL-pages are exhibited in Net 2, where we observe 24 blogs (out of 990) that have the most intensive ties with 11 firms (out of 18 firms). The meaning of this selection is that these 24 blogs engage in relatively more comparisons of firms facilitating the co-presence of these pharma companies in a common context. The blogs with the most intensive ties to multiple pharma context are – Pharmalot, Impactivity blog, and Canada’s shame.
Net 1. All ties between companies and page/URL

Net 2. More then 5 ties between companies and URL-pages (del pendants)
Net 3. Companies vs. key words in block A: HEALTH (normalised value)

Net 4. Interconnected key words in the semantic field of block A: HEALTH (normalised value)
3.2. Mapping of the key actors and the topics on which blog-participants publish

The key actors in our analysis were blogs and references in text to pharmaceutical companies or selected issues (key words). The mapping of key actors included analysis of the semantic associations between the pharma companies and the key-words that represent the semantic field of European pharmaceuticals and European healthcare – grouped as Health, Drugs, Disease, and Regulation. As we are searching for the co-presence of key-words in blog content, the relationships that we reveal are contextual and represent semantic associations.

Net 3 identifies how individual pharmaceutical companies are located in the semantic field that corresponds with European pharmaceuticals and health care. We have used normalised value to counteract the size effect where all firms are equal irrespective of the volume of blogs in which they have been referred to. In Net 3 Astrazeneca, Merck & Co, and Takeda appear most central to the debates surrounding healthcare issues, including medicine, diagnostics, and public health issues. Pfizer is mostly associated with generic categories such as healthcare, health care system, and health policy. Eli Lilly appears as an isolate in this map, which means that it has no preferential associations with any particular issues related to healthcare, but exhibits equal presence in all related discussions. This can be interpreted as ‘broad and/or indiscriminate impact’.

The overall structure of the semantic field reveals that some categories are engaged in complex structural relationships and represent the semantic core, while others are more peripheral and less influential. The semantic categories of diagnostics, public health, medicine, and health care services represent the core of this semantic field, and are instrumental for the discussions around the pharmaceutical companies.

The same semantic field represented in a one mode graph exhibits different relationships, where we observe three distinctive semantic components. The largest and most densely connected component comprises of categories that refer to restructuring of health care system and private health care. The semantic categories interconnected in this component represent associations that underpin the context of health discussions. Another large component comprises of four generic categories that have very high level of presence in the cumulative blog content, but remain as a separate semantic group, connected to the rest of the semantic field via healthcare trust and health care system. Obviously the blog discussions make multiple references to medicine, health care and patients. The third small semantic component comprises of an institutional representations such as hospital, European healthcare and healthcare business. Semantic bridges between these components are healthcare trust and diagnostics.
Each of these components contains associations that reflect patterns in the blog content and indicates information stream. This interpretation, however, requires more in-depth text analysis. We present here only one area selected for our semantic analysis – issues related to health.

3.3. Mapping of relationships between blogs

All blogs and URL-pages in our database are interconnected as they all represent one common semantic field – drawn by the use of key-words from the selection criteria. However, some of them are ‘more’ connected than others as they show awareness of each other and send links to each other. These links we have qualified as internal links as they point at connectivity within our blog selection in the database.

From the map on Net 5 we can conclude that the blogo-sphere in our field is very fragmented. There are only occasional links (cross-references) between URL-pages forming a large number of dyads (88%) and short tails. There is only one small group of URL-pages in the centre – a central interconnected component representing 18% of the total population, which is linked to a blog called ‘Garbage Garbage’, where interconnectedness emerges. Most importantly, this internal ‘visibility’.

3.4. Mapping of the impact

There are different ways for evaluating the potential impact of an actor. One of the established methods is evaluation of the centrality of blogs and URL-pages (i.e. how central and interconnected is each URL-page from a blog (Net 5). However, this analysis reveals very limited impact – as healthcare blogs are disconnected – informing only their specific audiences, where the audiences do not seem aware of other blog-audiences, i.e. making limited reference to other blogs with similar thematics. The number of dyadic links shows that there are only occasional links between blocks and URL-pages, but each blog entry exists mainly by itself.

4. Conclusions & managerial implications

The blog-space is a dynamic configuration of the internet with continuously changing entries and exits. The dynamics is exhibited by a discrepancy between registered new blogs (acknowledged as URL links) and available blogs (obtained from blog-servers). We also observed very little interaction and public response compared to the volume of text that is broadcasted in publicly accessible blogs. The majority of blog postings have not received any comments from the audience and this excluded the question about emerging online communities. If comments were posted, they often evolved in one thread that followed up upon one article. They were written within one to two days from the original post and did not engage in a serious discussion or other behavioural response – to confirm influence.

The structural relationships and positions of actors reveal only potential influence, and this is a confirm observation for all types of actors – pharma firms, or key blogs and semantic categories.

The high connectivity between firms indicates that most blogs compare across a wide set of companies, and this may be interpreted as information that is being generated by professional activists, media activists, or actors with a broad set of observations in our field.

The fragmented internal connectivity also indicates of insignificant impact from one blog to another blog’s audience. Such a fragmented intelligence system can not offer an effective means for information distribution.

There are many substantially different formats of blogs that are in use, and it seems that there is no dominant pattern of format emerging. Most blogs have options for enabling comments and other interactions, such as tagging or emailing an article. However, their classification as blogs and/ or their selection by blog-search engines is often due to technical features such as meta-tags in the HTML code of a web page.

The majority of blogs in the area of European Pharmaceuticals and European Healthcare are technical and organisational experimentations and explorations that aim to broadcast information. There are two main types of blog-news - generalist news (blogs established by the main media with publications or specialised sections on healthcare and pharmaceuticals), and specialised news (blogs established as specialised sources of information on medicine and healthcare). There are some ‘community-type’ blogs that stir community interactions, and some personal blogs – as individual attempts for expression of opinion. Almost all blogs have included in their registration entry a copyright claim.

Many blogs use automated facilities for organizing and structuring the information, e.g. via time-based archiving of posts and tag-based aggregation. The main volume of blogs has emerged from the beginning of 2008. Multiple blog search-engines are available. They vary substantially in the results and ranking of the results that they retrieve from the world-wide-web.

For the purpose of our comprehensive analysis we used two frameworks: one included the entire semantic field of European pharmaceuticals and European healthcare; and the other – the leading 19 pharmaceutical companies. All pharmaceutical companies included in our search have a presence in the blog-space – with exception of Lacer. Large firms attract a lot more attention, and the reference to Pfizer is dominant (289 URL-pages for the period up to July 2008), followed by GlaxoSmithKline (205 URL-pages), Novartis
The unique methodology that we used enabled us to retrieve information on blogs for blog ranking according to their importance in a selected semantic field. Our maps represent contextual graphs that describe location of URL pages, semantic categories, or firms in context. These maps can be used as guidelines in expertise seeking, or finding patterns in blogs’ evolution.

5. References


[7] Gamon, M., Basu, S., Belenko, D., Fisher, D., Hurst, M., Konig, A. (2008) BLEWS: Using Blogs to Provide more in-depth information on disease areas and methods of treatment. Both players attract fairly similar public attention in terms of comments and interactions, which is still very low (within 1-2 days only after the publication). The majority of blogs have some association with private organisations that manage the blogs, which suggests that serious and long-lasting blogs will exhibit the influence of some organisational strategies.

Due to the high volume of entries in the blog-space, research is recommended on a narrow set of categories to demarcate a narrow semantic space for blog search and for analysis. Our choice of the 19 pharmaceutical companies is a successful strategy, as it can draw clear boundaries for the population of URL-pages in the database.

The four semantic blocks that were identified in our search (DRUGS, DISEASE, INDUSTRY, and INSTITUTIONS) require independent in-depth research. Representative research on each semantic blog will reveal in-depth associations, meaning and values that underpin discussions in each semantic blog. Such results will have a direct use in marketing and public relations.

The unique methodology that we used enabled us to retrieve information on blogs for blog ranking according to their importance in a selected semantic field. Our maps represent contextual graphs that describe location of URL pages, semantic categories, or firms in context. These maps can be used as guidelines in expertise seeking, or finding patterns in blogs’ evolution.

5. References


[7] Gamon, M., Basu, S., Belenko, D., Fisher, D., Hurst, M., Konig, A. (2008) BLEWS: Using Blogs to Provide more in-depth information on disease areas and methods of treatment. Both players attract fairly similar public attention in terms of comments and interactions, which is still very low (within 1-2 days only after the publication). The majority of blogs have some association with private organisations that manage the blogs, which suggests that serious and long-lasting blogs will exhibit the influence of some organisational strategies.

Due to the high volume of entries in the blog-space, research is recommended on a narrow set of categories to demarcate a narrow semantic space for blog search and for analysis. Our choice of the 19 pharmaceutical companies is a successful strategy, as it can draw clear boundaries for the population of URL-pages in the database.

The four semantic blocks that were identified in our search (DRUGS, DISEASE, INDUSTRY, and INSTITUTIONS) require independent in-depth research. Representative research on each semantic blog will reveal in-depth associations, meaning and values that underpin discussions in each semantic blog. Such results will have a direct use in marketing and public relations.

The unique methodology that we used enabled us to retrieve information on blogs for blog ranking according to their importance in a selected semantic field. Our maps represent contextual graphs that describe location of URL pages, semantic categories, or firms in context. These maps can be used as guidelines in expertise seeking, or finding patterns in blogs’ evolution.


