The Studies of Blogs and Online Communities: From Information to Knowledge and Action

Emanuela Todeva and Donka Keskinova

Abstract This research addresses the question of whether the rise of blogs as a rich information source may create new opinion leaders that transform and challenge the traditionally held public views on drugs and European health care. We investigate blogs that discuss issues related to European health care and European pharmaceuticals for a selected 6 month period. In our approach of the blog space, we take a sociological perspective and design a multistage methodology for data collection and data analysis that differs from the traditionally used crawling techniques by computer scientists.

The results reveal that in spite of the high volume of blogs for the investigated period, only a small number are interlinked by mutual referrals. The emerging network configuration is represented by a small core component with a large number of dyads, or short tails, which represents a fragmented community space. Our content analysis reveals that the information broadcasted in blogs shows emerging semantic differentiation related to specific health issues and disease categories. Our findings support the conclusion that in spite of the high technical Internet connectivity facilitated by search engines and Internet crawling tools, community interaction is limited, and there is no evidence of online crowd or collective action.

1 Introduction

Blogs are related to a group of interactive technologies developed for computer-supported cooperative work in a network and in online virtual environments. These interactive Web 2.0 technologies represent tools for individual online publishing of

E. Todeva (✉)
University of Surrey, Surrey, UK

D. Keskinova
Plovdiv University, Plovdiv, Bulgaria

N. Agarwal et al. (eds.), *Online Collective Action*, Lecture Notes in Social Networks,
DOI 10.1007/978-3-7091-1340-0_6,

content, collaborative editing, collaborative learning, or other collaborative online activities. The analysis of these Internet technologies reveals that often the aim of the design (i.e. to facilitate collaboration and interaction) is substituted for the effect of the use of the technology (i.e. the collaboration itself). Such analysis infers social connectivity on the basis of enabled actors, rather than actors actively engaged in interaction. Technological capabilities are substituted with the actual use of the technology and the social processes at the human side of human–computer interaction. The impact of technical connectivity on knowledge creation and dissemination is sometimes analytically substituted for human interaction and community practice.

There is little attempt to discriminate between the use of an interactive technology for broadcasting information and the transformation of this public communication into interaction, information sharing, or a full-scale community collaboration. From a substantive point of view, the question that we address is at what point mass communication becomes internalised in personal lifestyles and when a group of communicating and interacting people becomes a community or when group communication, group awareness, and group coordination transform group interactions into community practices.

In our blog analysis, we aim to discriminate clearly between technical connectivity and human behaviour in an interconnected environment (i.e. interconnected computers and web pages vs. human interactions). Although blogs offer both access to content and a connectivity platform, the presence of technical connectivity should not be used as evidence of a social relationship or other association in an affiliation network. In the first part of this chapter, we develop a theoretical framework for analysis of three distinctive stages of blogging, i.e. information sharing, knowledge creation, and community action and interaction. Our theoretical discussion establishes the conceptual background to our research, looking at the distinction between online communication and online collaboration and interaction.

In order to explore empirically these issues, we build a database of blogs that discuss issues of European pharmaceuticals and European health care for a 6 month period (January–June 2008). Our database includes URL-titles, the referral links between individual blogs, and the semantic context of the blogs. We employ network analysis methods to determine the emergent connectivity.

The methodology for the empirical investigation is outlined in the second part of this chapter, and we label it as a social science methodology for blog analysis. The third part of the paper presents our empirical findings on the structure of the European Pharma blog space, the identification of key blogs that dominate the distribution of information, analysis of the relationship between blogs, evaluation of the semantic structure of the discussion on health issues and disease, and analysis of the referral connectivity in the blog-space.

2 Conceptual Background

Blogs are seen as a new publishing medium—both employed by and challenging the established mass media firms and practices (Gamon et al. 2008). Blogs are portrayed as an enabling technology for personalised expression of opinion and a medium enabling individuals to speak for themselves (Cardon et al. 2007). In all these cases, the expectations are that blogs can and do influence public opinion—either as an Internet broadcasting tool or as a medium used by key opinion leaders, by specialised interest groups, and in communities of practice to promote their activities (Magnus Berquist et al. 2003). As an enabling technology, blogs along with other web technologies facilitate social connectivity on the web, which leads to the evolution of a complex social topology of the web (Barabási et al. 2003).

There are many questions that need to be addressed in this context—both at theoretical and empirical levels. There are also many disciplines that are currently researching the Internet space populated by Web 2.0 users and creators of web content. Research of the web contributes to our knowledge and understanding of the processes that take place in blogging and the impact of various enabling Internet technologies.

Media studies have emphasised that information broadcasting is an effective tool for shaping public opinion and consumer preferences (Eisengerg et al. 1985). At the same time, communication studies confirm that communication relationships require exchange of information, which demonstrates shared meaning between senders and receivers and the creation of a common semantic field. Although communication relationships may take place outside communities, they can take place only within a particular semantic field, where the actors understand each other and share common meaning, semantic frames, and views of the world at large. Semantic fields hence represent the boundaries of distributed communication practices and knowledge and intelligence systems. Broadcasting and receiving information represents a communication relationship that has only one social component—common language code and the meaning of the transmitted content of information. Connectivity within semantic fields, on the other hand, can attribute social connectivity among actors who share a common language, meaning, and interests.

Social connectivity in this sense means social interactions based on the sent and received information and internalisation of this information. Evidence of such social connectivity is any thought or behaviour that is associated with received information. A crowd of people that receive the same information and 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 (i.e. a community). Social media can involve both broadcasting to a crowd and engaging online community in online information sharing. Both of these trigger collective action, although this is not necessarily equivalent to the received communication.

The literature on online communities has addressed the issues of the nature of virtual communities and the boundaries between communities of practice and

knowledge communities as distributed intelligence systems. There is also an effort to draw a distinction between social interactions, human–computer interactions, and online communications (Wellman 2001). Distributed intelligence systems are based on human–computer interactions, information storage and exchange, and online communication. Such a system, however, requires translation of information and social interactions that attribute meaning to the exchanged information in order to resemble a knowledge community within a shared semantic field.

Being part of the Internet, blogs by their nature allow interconnectivity and interaction, and there are expectations that they facilitate the formation of online communities (Cambrosio et al. 2006). The presence of social aggregation (i.e. interconnected social actors), 'long enough' connectivity between these actors, and the enactment of online mediated personal relationships (i.e. interactions and direct one-to-one communication) are seen as evidence of these online communities (Cambrosio et al. 2006). We can assume that online communities that are engaged in repetitive communication share common meaning and represent a knowledge community.

In a community of practice, however, we are looking for evidence of community participation 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 and common patterns of behaviour following community opinion leaders.

Co-authorship and co-editing in distributed intelligence systems are clear evidence of communication relationships as well as evidence of a community membership whereby the co-authors share a common semantic field and act according to community-shared rules and practices. The existing definitions of a community with reference to a group of individuals who share intent, belief, resources, preferences, or needs (Cambrosio et al. 2006) 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 (Wellman 2001), and personal statements can be an evidence of a community membership. Link analysis without text analysis can reveal unilateral or reciprocated communication relationships that can substantiate community interactions, but should not be subsumed as such. Feedback and response to communication resemble social interaction that can confirm a level of community affiliation and individual action. Even the most intensive participation in blogs may not be taken as evidence of collective action as it represents communication interaction and not necessarily behavioural contagion.

The typology of blogs circulated in the literature refers to a number of Internet-based shared contents such as online personal diaries, collections of links to other sites, or online public forums devoted to specific topics (Lin et al. 2007). These all are online broadcasting tools that disseminate information and constitute the infrastructure for the online communication process. In terms of their impact, there is a need not only to confirm that the broadcasted information has been viewed by some audience, but also that some form of interaction between a blogger as a sender and a blogger as a recipient of the information has taken place. Clear evidence of impact

and social interaction are comments made or acting under the influence of this information. 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 and acting under the influence of the group. This is referred to as 'other directed' or social contagion of behaviour in crowds (Russ 2007). Posting a comment can be considered as evidence of interaction and a 'community' relationship (Cambrosio et al. 2006). 'Html' referrals in blogs also can be interpreted as enactment of a relationship where the blogger who posts the referral link expresses some knowledge and attitude to the referred blog.

In terms of sources and effects of influence, we have to discriminate between a source of information (i.e. online publishing/broadcasting medium), a potential recipient of information (all Internet users that have online access to the source), and an interaction between the source and the recipient or a reaction by the recipient in response specifically to viewing the source (Todeva 2006a).

Communication in the Internet space is evidence of a membership in a distributed intelligence system, but is still not an evidence of a community relationship as it is not an evidence of shared meaning, sociability, and sense of belonging (Wellman 2001). In spite of the universal connectivity of the Internet, there is a difference between sending and receiving information, shared knowledge and meaning in a common semantic field (i.e. knowledge community), and two or more individuals acting in accord and agreement (i.e. community of practice). An observer to a community differs from a member of that community by the participation in coordinated/shared activities. Community membership for an aggregation of individuals can be attributed by 'sharing' particular beliefs and by enactment of these beliefs in coordinated practices.

The blogger who establishes the blog creates a 'community platform' where other individuals can join-in. A blogger that posts a comment on such a platform demonstrates a clear intent to join this community, and then it is in the realm of text exchanges and meaning sharing, where shared understanding and community relations emerge.

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 cannot 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 is necessary—to support a 'community' hypothesis for the blogosphere.

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 health care). 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 online discussions (i.e. knowledge communities and communities

of practice). In addition, we looked at semantic connectivity or dominant semantic relationships based on a selection of keywords that represent our semantic framework. Our approach to blog analysis and the methodology for the empirical investigation, described in the next section, have been informed mainly by social anthropology, communication theory, semiotics, and organisation theory.

3 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 World Wide Web (WWW) and to speculate on the 'potential' spread of influence across blogs linked by referral URLs (Leskovec et al. 2007). New algorithms for page ranking are among the issues that have attracted the attention of the computer scientists (Tseng et al. 2005; Esmaili et al. 2006; Kritikopoulos et al. 2007). Searching core social structures and cyber-communities on the web has led to the development of a number of mapping techniques identifying homogeneous groups of blogs by topic (Dourisboure et al. 2008). 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 (Adamic and Glance 2005) or music blogs (Cambrosio et al. 2006).

Our selection of the semantic framework identifies a segment of the blogosphere, which is characterised by an overlap of commercial and public interest. As such, it is expected that the content of the blogs will reflect both commercial and private views. Being a broadcasting media, it is expected also that the content will reflect current events and some deeper underlying individual predispositions to health care and to the pharma industry and products as well as self-expression and sharing of personal experience.

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 keywords identified as representative of our semantic framework.

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.). Blog analysis from a computer science perspective implies different techniques such as text, graph, or sequence mining (www 2008). We differentiate from these approaches by developing an alternative methodology for blog analysis that employs simultaneously content and relational analysis in order to evaluate the blog impact as emerging associations in a specified semantic field.

We have adopted a broad agency approach to the Internet where actors can be either pharmaceutical firms discussed in blogs or critical health care issues (expressed by keywords in content) or the blog URL pages themselves. Our application of network analysis of 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 or a knowledge community that shapes a common semantic field and initiates a community action.

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) (Cambrosio et al. 2006).

Our methodology for blog search and blog analysis comprises seven main stages, including building a comprehensive database with the full population of blogs that correspond to our selection criteria, cleaning of the database, evaluation of reference links and semantic associations within blog URLs, and mapping of semantic links and connectivity ties at micro (within blogs) and macro (across blogs) levels.

3.1 Development of the Selection Criteria

Our first step was to demarcate the boundaries of our semantic framework in order to explore it in detail. We conducted text analysis of the news on European health care and European pharmaceuticals broadcasted in official online media between January and June 2008 and identified the 'search keywords' reflecting key events and dominant issues during this period. We grouped these selection words in six distinctive semantic groups (health, drugs, diseases, industry, regulation, region). These groups of keywords demarcated the boundaries of our semantic framework within which we looked at blog referrals and semantic associations within blogs or dyads of keywords with relatively high co-presence in a URL page from the entire blog space.

3.2 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 was that it produced at the time of investigation the minimum duplications of URL pages on initial search, with the maximum of total URL pages identified in a filtered query.

3.3 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/health care)
- Geophysical relevance (Europe, UK/England, France, Germany, Spain)
- One of the selected keywords (we constrained the sample of keywords 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 has entries, which mentioned at least one company name, and the database itself has minimum semantic noise across the population of blogs.

3.4 Building the Database

The final database was generated as the total population of all obtainable blogs that were active at the time of the research and present on the internet for the selected period of 6 months and contained at least one combination of keywords from our search criteria. We downloaded full blog details of these blog URL pages.

3.5 Cleaning the Database

After filtering the majority of duplications by the search engine itself (the difference between visible and obtainable), blogs led to an automated reduction of 85 %, and we cleaned further the database at three additional stages (see Table 1). 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 (1) cleaning of duplicate URL pages; (2) cleaning of 'empty URL pages' with size <2 kb information; (3) cleaning of 'shell URL pages' that contain dictionaries, job announcements, lists of URLs without text, URL classifications, and 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 contained shared blogs referring to more than one company (keyword) and 633 blogs (or 64 %) as periphery—i.e. blogs related to only one company (keyword) (observed as pendants on Net 1).

Table 1 Population size

	Total available URLs	Total obtained	Less duplicates and 'shells'[a] (final population)	Representative pages
Total URLs	11,824	2,995	990	633

[a]Duplicate pages—URLs with the same size and content, and registered as different web-links. Shells—URLs that contain lists of words and/or URLs, without other meaningful content

3.6 Developing Blog Attributes (Primary Analysis)

Our primary analysis involved some preliminary observations and developing blog attributes through Internet count of key blog indicators. We calculated four additional indicators as blog attributes per each URL: *size of URL in kb, cross reference between URLs in DB* (as internal hyperlinks), *cross reference to other blogs* (number of external hyperlinks), and *number of occurrences of individual keywords per URL page* (based on the six identified semantic groups). Some of these indicators were used for additional filtering of the data, and the final numbers were recorded after the cleaning process was completed.

3.7 Data Analysis and Network Mapping

The secondary analysis of mapping the selected population of blogs involved mapping of commercial and private broadcasting of information by bloggers and the discussions that this information triggers. We focused on different types of actors that formed a heterogeneous information system. These actors were (1) the owners of the blogs (expressing opinion), (2) the blogs themselves (as cumulative content), and (3) the shared content (through comments, referrals, and embedded URLs). Our network mapping aimed to reveal (1) the emergent structural concentrations of different types of actors and their network position inferring potential information sharing and source of influence, (2) mapping of the content of the text and discussion of information and the semantic links between blogs using one-mode and two-mode graphs, and (3) mapping the emergence of a shared semantic field where we can observe evidence of interactions.

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 (Cambrosio et al. 2006). This method allows to analyse the co-occurrence of 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 of the heterogeneous network system, we constituted and interpreted the following 'relationships':

- Associations between blogs and pharmaceutical firms (Nets 1 and 2). The similarity measure in these graphs represents a co-occurrence of pharmaceutical firms in blog contents of individual blogs.
- Association between pharmaceutical firms and semantic categories in five groups (Health, Drugs, Disease, Industry, and Regulation). The similarity measures in these graphs represent significant ties based on a co-occurrence of keywords and names of pharmaceutical companies.
- Associations within semantic fields (Net 3 for Health and Net 4 for Disease). The similarity measures in these graphs represent a co-occurrence of keywords from the semantic groups on health and disease in blog contents of individual blogs.
- Associations and cross reference between blogs (internal links between URL pages in a database) (Net 5). The similarity measure in this graph represents cross references between blogs.

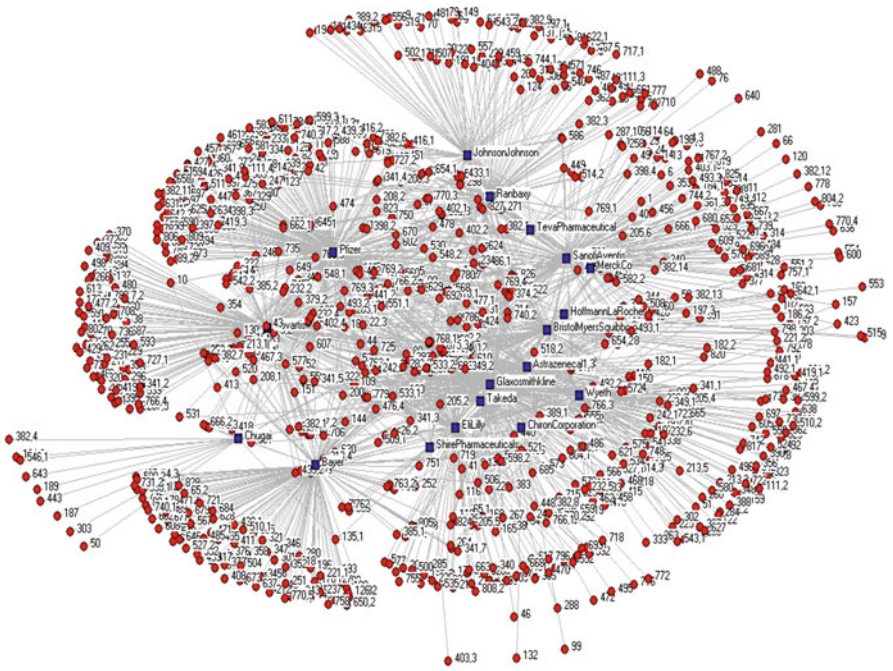
For Nets 1, 2, and 5, we have used the absolute value of ties, and for Nets 3 and 4, we have used values based on standardised residuals (Chi-square) (Todeva 2006b).

4 Overview of Results

The results from the blog analysis are grouped in three main sections (1) mapping of the blog space of our selected framework and mapping the key actors in this space as well as pathways of information dissemination; (2) mapping of the topics on which blog participants publish content (semantic analysis of emerging associations within a knowledge community); (3) mapping of relationships between blogs—comments, feedback, group communications, and community interactions.

4.1 Mapping of the Information Dissemination within the Blog Space

The first two maps show a distribution of URL pages and their association with a particular pharmaceutical company. Net 1 shows that Pfizer, Bayer, and Novartis exhibit unique profiles with their clouds of pendants or URLs dedicated to one company only. These three companies along with two others (Sanofi Aventis and GlaxoSmithKline) form a group of similar firms that are the most referred to in the blog space. These firms have the largest numbers of unique URL pages referring only to one of them (between 50 and 96 referral URLs) and the largest number of URLs that compare two or more of them (between 80 and 183 shared pages for each company). Unique URLs that refer to one company only are graphically presented as pendants on Net 1, and we have labelled them '*representative blogs*', while URLs that refer to two or more firms are labelled '*comparative blogs*'.

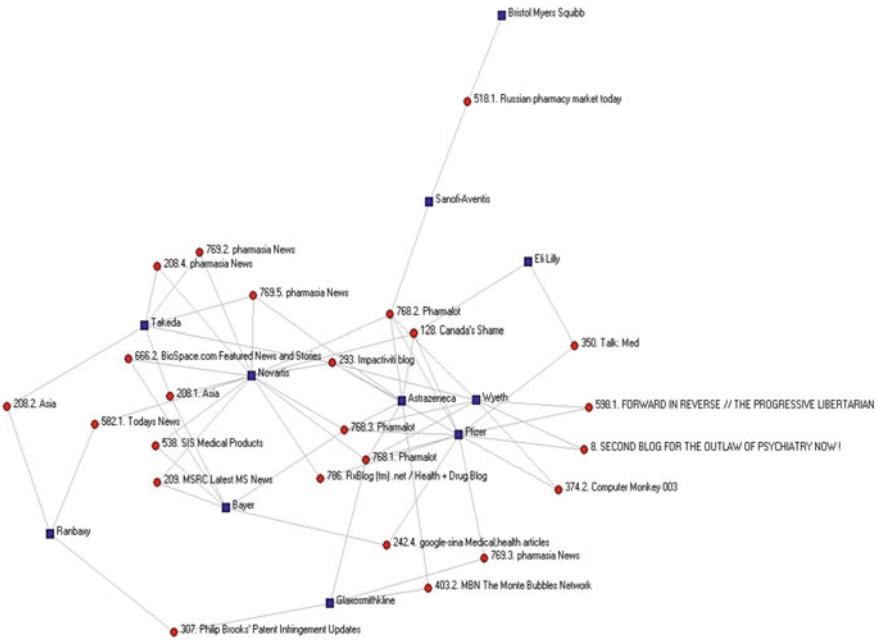


Net 1 All ties between companies and URL-pages

Representative blogs deliver information to the audience regarding one firm only, while comparative blogs enable bloggers to compare and contrast two or more firms.

Among the comparative blogs, we can also discriminate between dyadic and multilateral comparisons that enable bloggers to compare and contrast information on multiple firms. Using the degree centrality measures (DC), we can identify that three of the companies in this group stand out as the most discussed companies with the most similar referral profiles. These are Pfizer (DC = 611), GlaxoSmithKline (DC = 556), and Novartis (DC = 483). The graph in Net 2 displays that while Pfizer and Novartis maintain strong connections with comparative blogs, GSK has significantly low strong ties with blogs (only four connections with comparative blogs).

The second group of seven firms in Net 1 (Merck & Co., Chugai, Takeda, Teva Pharmaceutical, Hoffmann La Roche, Chiron Corporation, Shire Pharmaceuticals) represents an opposite type to the first one—with a minimum number of single referral pages (between 1 and 11) and a fairly low presence in a comparative and competitive context (between 4 and 50 shared URL pages for each company) (Net 1). These firms have a fairly low profile in our population of blogs, which means that bloggers have received information from significantly low Internet sources.



Net 2 More than five ties between companies and URL-pages *(del pendants)*

The third group of six firms (Ranbaxy, Eli Lilly, Wyeth, Johnson & Johnson, Astra Zeneca, Bristol Myers Squibb) stand between the two already described groups, with representative blogs between 16 and 40 and comparative blogs between 37 and 105). These results confirm that the popularity of firms is associated with both unique promotion strategies from representative blogs and online discussions in a comparative and competitive context through comparative blogs. The information impact hence is equally sensitive to volume and strategic approach. The number of comparative blogs is nearly twice bigger (633 blogs) than the number of representative blogs (357 blogs) (Table 1).

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 across 11 firms (out of 18 firms). The interpretation of this map is that these 24 blogs engage in the most intensive comparisons of firms facilitating semantic associations between health issues, pharma companies, and other disease and drugs semantic categories by their co-presence in a common context. Blogs with the most intensive ties to multiple pharma contexts are Pharmalot, Impactivity blog, and Canada's shame.

From the map on Net 2, we can identify two different groups of firms that are most directly compared. These are Pfizer, Astra Zeneca, and Weyth and Beyer, Takeda, and Novartis. The second group of firms mainly emerges through

specialised discussions in a few blogs (i.e. blog Asia) that hold strong connections and comparisons.

In terms of scope and reach of the information dissemination, most blogs in our selection have a large number of URL referrals, where 36 % of our population have between 21 and 50 blog referrals, 25 % have between 51 and 100 referrals, 19 % have between 101 and 200 referrals, and 16 % have more than 201 referrals. These numbers indicate a rich information environment where further text analysis may reveal the context for these comparisons.

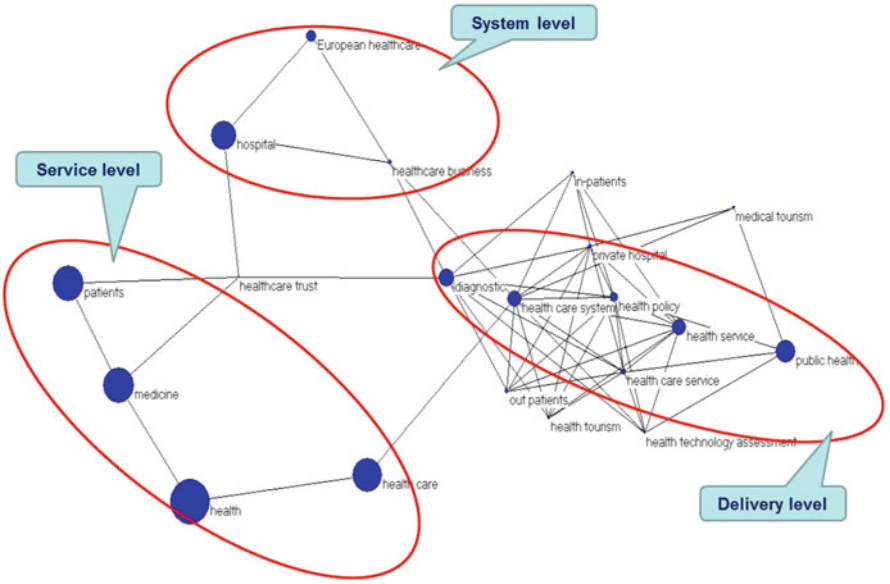
4.2 *Mapping the Semantic Association and Knowledge Creation in Blogs*

The concepts that emerge as core categories in our semantic framework are *patient*, *hospital*, *disease,* and *medicine*. Our network analysis of this broad semantic context identifies an additional set of keywords that focus the content of the discussion around *health care*, *public health*, *community*, *risk*, *regulation*, as well as the two most popular diseases—*cancer* and *diabetes* (Nets 3 and 4).

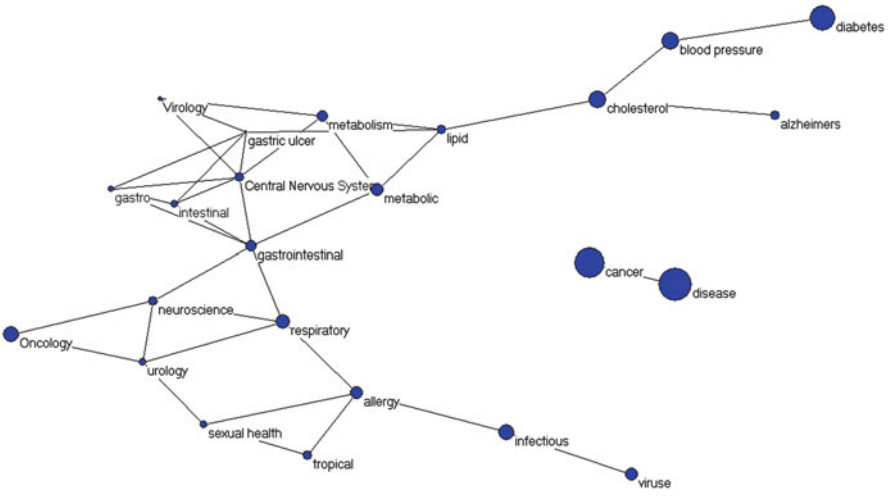
The key actors in our semantic analysis are references in text to pharmaceutical companies or to selected keywords grouped in the semantic subfields of *Health*, *Drugs*, *Disease*, *Industry,* and *Regulation*. In our semantic mapping, we have searched for the co-presence of keywords in blog content and for semantic associations that reveal context and represent knowledge structures. We undertook a semantic analysis of each semantic subfield independently. We present here the results from the semantic analysis of two semantic subfields—*Health* (Net 3) and *Disease* (Net 4).

In our semantic analysis of the first subfield *Health,* we reveal that Astrazeneca, Merck & Co., and Takeda appear most central to the debates surrounding health care issues, including the categories of *medicine*, *diagnostics,* and *public health*. Pfizer is mostly associated with generic categories such as *health care*, *health care system*, and *health policy*. Eli Lily appears as an isolate in this map, which means that it has no preferential associations with any particular issues related to health care, but exhibits equal presence in all related discussions. This can be interpreted as 'broad and/or indiscriminate impact'.

The overall structure of the semantic subfield *Health* reveals that some categories are engaged in complex semantic 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 most dominant categories are *health*, which occurs in 70 % of the URL pages, followed by *medicine* (50 %), *patients* (47 %), and *health care* (44 %).



Net 3 Interconnected key words in the semantic field of HEALTH *(normalised value)*



Net 4 Interconnected key words in the semantic field of DISEASE *(normalised value)*

The same semantic subfield *Health* represented in a one-mode graph (Net 3) exhibits different relationships where we observe three distinctive semantic components. These refer to concepts at **service level** (medicine, patients, trust), at **delivery level** (private hospital, outpatients, health tourism, health technology), and at **system level** (European health care, health care business, hospital).

The largest and most densely connected component comprises categories that refer to the delivery system or the restructuring of health care and private health care. The semantic categories interconnected in this component represent associations that underpin the core context of discussions on health. Each of the components contains associations that reflect patterns in the blog content and indicate knowledge structures that require more in-depth text analysis.

The semantic analysis of the subfield *Disease* shows the dominant position of the concept of *cancer* (in 36 % of the URLs) and its association with the generic category of *disease* (in 46 % of the URLs). This semantic association stands alone compared to all other semantic categories. Both concepts of *cancer* and *disease* have the highest occurrence in URL pages (359 and 459 pages respectively) and the highest co-occurrence in the same text. The rest of the categories in the subfield of *Disease* resemble a large interconnected component, bridged by the disease categories of *allergy* (6 % of URLs), *respiratory* (9 %), *gastrointestinal* (5 %), *metabolic* (6 %), and *cholesterol* (13 %). These bridges contain generic disease categories that have a moderate occurrence in the total population of URLs, but have high co-occurrence in the 'disease' subfield (Net 4). All categories in the large component on Net 4 exhibit more rich contexts of the discussions compared with the discussions of *cancer* and *disease*.

4.3 *Mapping of the Impact of Information Dissemination and the Emergence of Community Relationships*

There are different ways for evaluating the potential impact of an actor. Although communication theory acknowledges that the impact of information dissemination depends on the coding (at the point of the sender—i.e. language, implied meaning, and structure of the message), it also depends on the noise during transmission and the decoding at the point of the receiver. High centrality of senders assures that in their broadcasting role they have high engagement with receivers. The impact of the information, however, has to be sought at the receiver's end.

One of the established methods for evaluation of the impact of blogs and URL pages is how central and interconnected is each URL page. Our analysis in Net 5 reveals very limited impact and connectivity. Our selection of health care blogs shows that out of the 990 URL pages, only 112 are connected, forming 40 disconnected components, 33 of which are dyads, 3 are open triads, and only 4 blogs resemble some form of connectivity with closure and mutual referrals. Most of the blogs are informing only their specific audiences, where the audiences do not seem aware of other blog audiences, i.e. making limited references to other blogs with similar thematic discussion. Most of the external links of our selection of blogs are to URL pages that have more generic content and are not included in our selection.

Nearly 16 % of the blogs in our selection have more than 200 external links or referrals to other blogs, and approximately 20 % have between 101 and 200 external

Net 5 Ties between URL-pages based of internal links [*node-size is equivalent to the size of the blog (kb)*]

links. These numbers demonstrate that there are significant efforts to generate internal connectivity, which, however, remains low. The number of dyadic links (33) shows that there are only occasional links between URL pages, and this can be interpreted as evidence of emergent connectivity in the European health care blog space. The majority of blog entries, however, exist mainly by themselves (878 unconnected URL pages). The different size of the nodes in Net 5 is an evidence of some impact associated with a larger volume of information. The blogs with the largest size are Brand Pile, Guiltner Review, and Prather blog.

From the map on Net 5, we can conclude that the blogosphere in our field is very fragmented. There are only occasional links (cross references) between URL pages forming some dyads and short tails among 12 of the blogs in our population (only 12 connected URLs from the total population of 990). There is only one significant component of URL pages in the centre comprising of 21 interconnected URLs linked to a blog called 'Garbage Garbage', where interconnectedness emerges. Such a structural position represents internal 'visibility' for the content in these blogs, where further analysis may reveal the emergence of some interaction patterns. Text and discourse analysis, however, of this interconnected component was beyond our research scope, and, hence, we are not able to draw conclusions on which discussions and semantic associations are linked with this centrality and connectivity.

The final stage of our analysis focused on the question—to what extent the emerging connectivity in the 12 % of the blog space influences the semantic relations and associations within the knowledge field or the impact on the knowledge community of all 990 blogs (Table 1). We compared the results exhibited in Nets 3, 4, and 5 using our measure of internal connectivity (i.e. URL referrals to other blogs internally). We identified that in most cases there are insignificant differences between semantic connectivity within the interconnected 12 % of blogs and the semantic connectivity within the rest of the dispersed population (878 unconnected blogs). The exceptions from this pattern are the categories *diabetes*, *oncology*, *cancer,* and *blood pressure,* where the co-occurrence within the knowledge field of interconnected blogs is significantly higher than the use of these concepts in the disconnected population of blogs. The interpretation of these results is that these semantic categories represent stronger public interest that generates connectivity.

Other exceptions are the categories *metabolic* and *respiratory*, where the co-occurrence in the disconnected blogs is significantly higher than in the interconnected core. These are medical categories that have stronger potential impact in their generic use as medical terms rather than in the context of specific semantic associations used by blog users within the emergent knowledge community. Our emergent knowledge community of 112 blogs shares stronger meaning for the first set of categories—*diabetes*, *oncology*, *cancer,* and *blood pressure*, where acute public interest generates some form of connectivity in the form of cross references.

5 Conclusions and 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 in the total number of blog searches) and available blogs (blogs obtained for viewing). This is evidence of the noise associated with web crawling and simple computer-based techniques for mapping of the Internet.

The main volume of blogs related to our semantic framework has emerged from the beginning of 2007, and the volume of URL pages has grown rapidly. At the time of this research, it exhibits a dynamic public space where new stories appear continuously, shifting the attention to specific issues. Monitoring this space is essential for tracking major shifts in public opinion, although there is a problem to attribute direct authorship to individual blogs and to measure viewing impact.

Many blogs use automated facilities for organising and structuring the information, e.g. via time-based archiving of posts and tag-based aggregation. These facilities help to establish the semantic frame of individual blogs and facilitate the organisation and distribution of information. The results show emergent

semantic frames that can be interpreted as shared communication environments to support online community structure.

There are two main types of blog news—*generalist* news (blogs established by the main media with publications or specialised sections on health care and pharmaceuticals) and *specialised* news (blogs established as specialised sources of information on disease areas, methods of treatment, medicine, and health care). Both types attract fairly similar public attention in terms of comments and interactions, which is still very low (within 1–2 days after the publication). There are also some specific 'community-type' blogs that stir community interactions and some personal blogs—as personal diaries and individual attempts for expression of opinion. We observed very little interaction and public response compared to the volume of text that is broadcasted in publicly accessible blogs. The lack of a significant number of comments can be interpreted as lack of collective action and even as lack of individual action. This is contrary to current observations of oversubscribed social media such as U-Tube and eBay, where we can observe a distinctive volume of individual actions and events of crowd behaviour.

The typology of blogs above suggests that the system for distributed intelligence is not homogeneous and, hence, we may expect different structuring processes to take place. Almost all blogs have included in their registration entry a copyright claim. 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 institutional strategies and organisational arrangements (McGlohon et al. 2007) that are currently managing this Internet space.

The majority of blog postings represent online broadcasting as they do not receive any comments from the audience, and this excludes the question about emerging online communities. If comments are posted, they often evolve in one thread that follows up upon one article. They were written within 1 to 2 days from the original post and do not engage in a serious discussion or other behavioural response—in order to confirm influence. The majority of blogs in our selected framework can be seen as technical, personal, and organisational experimentations and explorations that aim to broadcast information about drugs, disease, and broad health care issues.

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 the web page.

We have attempted to reveal associations between blogs on the basis of external links (html references to the web pages), internal citation and referencing, and commonality of interests (addressing the same category from our semantic frame).

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 frame 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.

For the purpose of our comprehensive analysis, we used two approaches: One included the entire semantic framework, and the other—the leading 19 pharmaceutical firms and keywords in our semantic subfields for health and disease. All pharmaceutical companies included in our search have a presence in the blog space—with the 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 (194), Bayer (159), Sanofi Aventis (156), Eli Lilly (146), and the rest.

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

The fragmented internal connectivity between blogs, on the other hand, indicates limited impact from one blog to another blog's audience. Such a fragmented intelligence system cannot offer an effective means for information distribution and its integration as a knowledge community. In addition, the lack of interaction and commenting suggests limited impact and lack of social contagion.

The structural relationships and positions of actors reveal only potential influence, and this is a confirmed observation for all types of actors—for pharma firms, for key blogs, or for semantic categories. In answer to the question '*Could an anonymous blogger have as much influence over public debate as a recognised scientific expert?*', we can answer that although the broadcasting of opinion on the Internet enables access to a global audience, there is no evidence that this audience has been influenced or even engaged at a communication level.

The mapping of the entire blog space for European pharmaceuticals and European health care has a vaguely connected core of 36 % shared URL pages and a large periphery of single URL pages related to a single pharmaceutical company. From the periphery, one third (or 86 URL pages) refer only to Pfizer and to no other pharmaceutical firm from our selection. Although Pfizer appears to occupy a space fairly at a distance from other pharmaceutical firms, it is also directly compared with Johnson & Johnson, Merck & Co., and Takeda, particularly on issues related to *health care*, *health care system*, and *health policy*. In addition, Pfizer appears to be strongly connected to blogs such as Pharmalot, Pharmasia news, RxBlog, Talk: Med, Canada's shame, Computer monkey, Forward in reverse, and Google-Sina medical health—among others. The names of these blogs indicate a very broad information-broadcasting platform that may explain the lack of interaction.

Although Pfizer is a dominant actor in terms of volume of blogs in which it appears in reference, it does not appear to have a distinctive profile. It appears rather generalist—in the semantic subfields of *Drugs*, *Health,* and *Disease*. This is in contrast to some other pharmaceutical companies that appear closely associated with a particular treatment area or health issues. A more focused discussion can potentially engage bloggers more effectively, particularly if it involves sharing information on personal practice. The lack of such action categories among our semantic categories demonstrates that the online broadcasting of information does not trigger immediate action.

The relationships between URL pages in the blog space are still rare. One of the blogs that has created fairly dense internal and external links is Garbage-garbage, and additional research reveals that few months later it was no longer active. In this context, although we recommend further research on the content of individual blogs with higher centrality, we also acknowledge that these might be texts and information that may disappear from the public space as spontaneously as they had been posted.

The analysis of the semantic subfields of *Health* and *Disease* reveals emerging threads of inter-related issues as well as semantic distances. Among these patterns are close proximity between *medicine*, *patients,* and *health care trust*—on the one hand—and *private hospitals*, *medical tourism*, and *health care system*—on the other. These examples indicate emerging discussions in the online public domain.

Disease areas such as cancer, diabetes, and blood pressure are domineering by themselves, and other disease areas such as metabolic, gastrointestinal, and respiratory appear quite interconnected in the public domain. Large pharmaceutical companies appear to have a broader impact on the blog space dominating the context in articles and publications, while small firms appear most often in the shadow of another large pharmaceutical firm. In this context, Pfizer's association with Teva Pharmaceuticals or Wyeth is visible.

The semantic groups of keywords that were identified in our search (health, disease, drugs, industry, and regulation) require independent semantic analysis. Representative research of each semantic subfield is expected to reveal in-depth associations and meaning. Such results will have a direct use in marketing and public relations. Although this semantic analysis reveals connectivity within the blog space, it can be used in analysis of communication flows in other social media.

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 locations 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. The unique power of network mapping enables to bridge the gap between micro and macro from individual actors to macro representations of affiliation networks.

Acknowledgements Special thanks and acknowledgement for the contribution of a number of colleagues that actively helped with the finance of the empirical investigation, the development of the methodology, and the analysis of the data (David Parry, Chris Shilling, Hristo Karapchanski, Jana Diesner).

References

Adamic L, Glance N (2005) The political blogosphere and the 2004 U.S. election: Divided they blog, XIVth international world wide web conference, Chiba, Japan

Barabási A, Dezső Z, Ravasz E, Yook S, Oltvai Z (2003) Scale-free and hierarchical structures in complex networks. In: American Institute of Physic Conference Proceedings, 661(1)

Cambrosio A, Keating P, Mogoutov A (2006) Mapping the emergence and development of translational cancer research. Eur J Cancer 42(18):3140–3148

Cardon D, Delaunay-Teterel H, Fluckiger C, Prieur C (2007) Sociological typology of personal blogs, ICWSM'2007. International conference on weblogs and social media, Boulder, CO

Dourisboure Y, Geraci F, Pellegrini M (2008) Extraction and classification of dense communities in the web. XVIth international world wide web conference, Banff, AB

Eisengerg E, Farace R, Monge P, Bettinghaus E, Kurchner-Hawkins R, Miller K, Rothman L (1985) Communication linkages in inter-organisational systems. In: Dervin B, Voight M (eds) Progress in communication sciences, vol 6. Ablex, Norwood, NJ, pp 210–266

Esmaili K, Jamali M, Neshati M, Abolhassani H, Soltan-Zadeh Y (2006) Experiments on persian weblogs. XVth international world wide web conference, Edinburgh

Gamon M, Basu S, Belenko D, Fisher D, Hurst M, Konig A (2008) BLEWS: using blogs to provide context for news articles. Association for the Advancement of Artificial Intelligence

Kritikopoulos A, Sideri M, Varlamis I (2007) Blogrank: ranking on the blogosphere, ICWSM'2007, international conference on weblogs & social media, Boulder, CO

Leskovec J, McGlohon M, Faloutsos C, Glance N, Hurst M (2007) Cascading behaviour in large blog graphs. SIAM Data Mining

Lin J, Halavais A, Zhang B (2007) The blog network in America: blogs as indicators of relationships among US cities. Connections 28(2):22–30

Magnus Berquist M, Feller J, Ljungberg J (eds) (2003) Open source software movements and communities. In: Proceedings of the international conference on communities and technologies, Amsterdam, Netherlands, September, 2003

McGlohon M, Leskovec J, Faloutsos C, Hurst M, Glance N (2007) Finding patterns in blog shapes and blog evolution. ICWSM'2007, International conference on weblogs and social media, Boulder, CO

Russ C (2007) Online crowds—extraordinary mass behavior on the internet. In: Proceedings of I-MEDIA '07 and I-SEMANTICS '07, Graz, Austria, September 5–7, 2007

Todeva E (2006a) Business networks: strategy and structure. Taylor & Francis, New York, NY

Todeva E (2006b) Clusters in the south east of England. University of Surrey, Surrey

Tseng B, Tatemura J, Wu Y (2005) Tomographic clustering to visualize blog communities as mountain views. XIVth international world wide web conference, Chiba, Japan

Wellman B (2001) Computers as social networks. Science 293(5538):2031

www (2008) An introduction to web mining. http://www2008.org/program/program-tutorials-TF3.html