A practical approach to exploit public data available on the Internet to study healthcare issues

Dárlinton Carvalho[1], Wilma Madeira[2], Mirna Okamura[2], Carlos Lucena[1], Sérgio Zanetta[2]

[1]Department of Computer Science
Pontifical Catholic University of Rio de Janeiro (PUC-Rio)
Rua Marquês de São Vicente 225, Rio de Janeiro, Brazil

[2]Instituto de Responsabilidade Social Sírio-Libanês
Rua Peixoto Gomide 337, Bela Vista, São Paulo, Brazil

{dcarvalho,lucena}@inf.puc-rio.br

{wilma.madeira,mirna.okamura,sergio.zanetta}@hsl.org.br

Abstract. *There is an abundance of data and services publicly available on the Internet, enabling researchers to study society in new ways. However, it is challenging to conduct studies using these resources, prompting questions such as where to find good data sources and how to apply them to the study of healthcare issues. This paper introduces a practical approach for undertaking such studies. This approach has three stages: the first provides a big picture of searchable information on the Web; the second focuses on online communities, looking for discussions around topics of interest; and the last investigates deeply a specific online community reality, searching for answers to research questions. As an application example, a study is presented with results on drug abuse in Brazil.*

Resumo. *A abundância de dados e serviços disponíveis na Internet permite novos modos na realização de estudos sociais. Entretanto, usar esses recursos ainda é desafiador, levantando questões como onde encontrar boas fontes de dados e como utilizá-los em estudos sobre questões relevantes de saúde. Este artigo apresenta uma abordagem prática para realizar esses estudos. Essa abordagem tem três etapas, em que a primeira provê uma visão geral das informações disponíveis na Web, a segunda vai em direção ao universo das comunidades online, e a última aprofunda na análise de uma comunidade online buscando respostas para questões de pesquisa. Como um exemplo de aplicação, são apresentados resultados de um estudo sobre o uso de drogas no Brasil.*

1. Introduction

The Internet is a powerful communication tool. As a digital medium, the Internet has interesting peculiarities, for example the footprints left by human interactions on the Web like surfing, posts on social networking sites and discussions archived on online communities. There is a great debate about privacy issues and ethical standards regarding tracking users' behavior on the Internet as well as the use of such data [Whitehead 2007, Kozinets 2009, Eysenbach 2009]. However, there are also reliable sources of public data available on the Internet available for use in academic research (e.g. [Eysenbach 2009]). This paper introduces a practical approach for using valuable social data to conduct social studies in order to increase the comprehension of important healthcare issues (e.g. [Eysenbach 2009, Kozinets 2009, Greene et al. 2011, Bender et al. 2011, Madeira 2011]). This approach was designed with three stages to guide researchers through their objectives. It starts with a broad content analysis, which aims to survey the whole Web, and goes on searching for a specific online community in social networking sites, until finding one that looks promising with regards to the research questions. It is a systematic use of free data available on the Web supported by computational tools, and therefore a guide to researchers.

The initial approach is similar to that of someone with a medical condition; it uses search engines. It is increasingly popular to search on the Web about symptoms and medication, often before looking for a doctor. Looking at the way people search for information on the Web can provide insightful data regarding population behavior [Eysenbach 2009]. Moreover, search engine companies, such as Google, provide a set of tools to help service providers create better ad campaigns (e.g. Google Insight for Searches, Google Trends, Yahoo Clues), as this is their main source of income. The idea is to take advantage of these systems and tools, looking for general population search trends, keywords and insights.

The second stage of the approach examines the world of online communities in social networking sites, looking for discussions about the topics of interest and identifying a panorama. Online communities are an important place for people on the Internet [Preece and Maloney-Krichmar 2005], where users can interact with others who share a common interest, exchanging information and finding support. A more recent phenomenon is the increasing use of social networking sites [Boyd and Ellison 2007], which also support the establishment of online communities. The approach is to use the social network sites' abilities to search for keywords related to the studying, identifying an overview of online communities.

The last stage of the approach is a deep analysis of one online community; searching to explain the community reality as related to the research questions. It should be a comprehensive study of the discussion presented in the community forum. The analysis employs social science and computational techniques. A valuable technique that can be used to describe the population is the Discourse of the Collective Subject [Lefevre and Lefevre 2006], a qualitative technique with roots in the Theory of Social Representations. Additional data available in the social networking site can also be automatically analyzed, generating big data aggregates as results.

Lastly, this research tries to bring together diferent disciplines, as is required to conduct such studies, and entering the arena of computational social science [Lazer et al. 2009]. The rest of the paper is organized as follows. Section 2 outlines the motivation for building the approach. The approach is presented in Section 3 followed by an example of its application, which is demonstrated in the Section 4. A discussion about this approach is presented in Section 5. Section 6 concludes this paper.

2. Social media data

The use of modern communication technology like the Internet introduces a new perspective for the study of society [Lazer et al. 2009]. In light of this paradigm shift, this paper presents an approach to exploit public content available on the Internet. This section presents the epistemology factored into the development of the approach.

The Internet is a medium used by a significant fraction of the Brazil population. According to comScore estimates from 2010 on Brazilian online audience [Banks 2011], there are around 77 million Internet users in Brazil, representing access from home, work and public locations (lan house, school, etc). This represents almost 40% of the Brazilian population [IBGE 2011] and the numbers are growing fast. Social network sites reach around 85% of Internet users in Brazil, according to the comScore statistics [Banks 2011]. Specially in the case of socially stigmatized illness, users go to the Internet to look for information and to share their experiences [Berger et al. 2005]. In online forums, individuals feel comfortable to talk about health conditions that would cause social stigma and embarrassment in face-to-face conversations. Anonymous interaction is a way that people can talk openly about their lives without embarrassment. Furthermore, besides exchanging information, members start to support others in the community, creating strong relationship ties.

The Internet usage leaves records that can be used to investigate health care issues [Eysenbach 2009]. The diversity of available data and research possibilities can be explored through different methods and techniques, Netnography [Kozinets 2009] being an inspiring one. The next subsection describes the realm of the Web as a data source, showing some applications, and presenting studies analyzing available data. In terms of healthcare investigation, the world of online communities is an outstanding place to find information about healthcare issues. Accordingly, a discussion on online community is presented based on studies highlighting its relevance to the study of healthcare issues.

2.1. Web as a data source in health

According to the Brazilian Internet Steering Committee [CGI 2011], 87% of users use the Internet to search for information and utilize online services. From these users, 35% use the Internet to search information related to healthcare or health services. In developed countries like USA and UK, the majority of users perform searches on healthcare topics. According to research of the Centers for Disease Control and Prevention [Cohen and Stussman 2010] from January through June 2009, 51% of American adults aged 18-64 had used the Internet to look up health information during the past 12 months. In the UK figures seem to be even bigger [Telegraph 2010]. A Porter Novelli EuroPNStyles survey showed that 65% of those questioned chose to search on the Internet when they want to know the answer to a medical query, compared to 43% who asked their doctor, while only 27% look for information via television shows. A mere 14% of interviewed people rely on government health information services.

The analysis of search, communication and publication behavior on the Internet can reveal interesting patterns about public health [Eysenbach 2009]. An example of this data application is the Google Flu Trends service[1]. The Google Flu Trends is a project that aims to detect and anticipate flu epidemics based on the analysis of search terms used on Google. Due to the population's increasing habit to search on the Web about symptoms and medication, even before looking for a doctor, Google can perform this real time Flu tracking system. According to Carneiro and Mylonakis [Carneiro and Mylonakis 2009],

[1] http://www.google.org/flutrends/

"Google Flu Trends can detect regional outbreaks of influenza 7–10 days before conventional Centers for Disease Control and Prevention surveillance systems."

Another remarkable example in this context is a Dengue surveillance system[2] that shows the evolution of dengue situations reported in Twitter. Dengue is a mosquito-borne infectious disease and a leading cause of illness and death in tropical and subtropical regions, including Brazil. The system was built based on an active surveillance methodology. Gomide et. al. [Gomide et al. 2011] found a high correlation between the number of cases reported by official statistics and the number of tweets posted during a same period.

2.2. Online Communities

Online communities are the most natural virtual habitat for people gathering to discuss their interests. The definition and delimitation of these virtual communities varies among the scientific disciplines and applications [Preece and Maloney-Krichmar 2005]. The broadest definitions see online community as any group of people on Internet. A stricter definition is that an online community must be lively, have a minimum number of members, policies, and purposes, and happen on the Internet. In this work, the investigation goes from a broad view of online communities towards a special online community, where answers to the research questions of interest are more likely to arise.

Influenced by ethnographic principles, Kozinets [Kozinets 2009] says that the Netnography observation happens at users' natural habitat. The content is detailed and contextualized in the community, and can be retrieved in a non-intrusive way, enabling an opportune, effective and efficient way to process it. In this way, users are not summoned to participate in a reactive fashion (e.g. online surveys), which enables the analysis of freely constructed opinions and manifestations.

A more recent phenomenon on the Internet is the emergence of social networking sites [Boyd and Ellison 2007], which have effectively served to recruit many users to the Web. They are a fertile environment for the establishment of online communities. In the context of health, these systems have been explored in many studies. Gold et. al. [Gold et al. 2011] realized a systematic examination of the use of online social networking sites for sexual health promotion; they found 178 sexual health promotion activities (meeting their inclusion criteria) and only one of these activities was identified in the literature. The authors comment, "SNSs are being used for sexual health promotion, although the extent to which they are utilised varies greatly, and the vast majority of activities are unreported in the scientific literature." Greene et. al. [Greene et al. 2011] performed a qualitative evaluation of communication by patients with diabetes with Facebook. For instance, they found out that "approximately two-thirds of posts included unsolicited sharing of diabetes management strategies, over 13% of posts provided specific feedback to information requested by other users, and almost 29% of posts featured an effort by the poster to provide emotional support to others as members of a community." Bender, Jimenez-Marroquin, and Jadad [Bender et al. 2011] analyzed the content of Breast Cancer Groups on Facebook, finding 620 breast cancer groups containing a total of 1,090,397 members. The groups were created for fundraising (277/620, 44.7%), awareness (236, 38.1%), product or service promotion related to fundraising or awareness (61, 9%), or patient/caregiver support (46, 7%). Lasker,

[2] http://www.observatorio.inweb.org.br/dengue/

Sogolow, and Sharim [Lasker et al. 2005] studied the role of an online community for people with primary biliary cirrhosis through the content analysis of a mailing list. Madeira [Madeira 2011] investigated online communities to describe transformations in the power relationship between physician and patient for her PhD Thesis. Whitehead [Whitehead 2007] provides an integrated review of the literature on methodological and ethical issues in Internet-mediated research in the field of health. There are many promising opportunities for study in the field of health, and the approach proposed in this work aims to guide researchers in performing these studies via the systematic use of free public data available on the Web.

3. A practical approach to go from data to knowledge

The proposed approach is a means by which to make opportunistic use of public data available on the Web in order to conduct research about healthcare issues. To accomplish this, the approach considers the use of several tools already available on the Web, as well as others tailored for specific purposes. This section presents some of the tools that can be used to acquire data for analysis and describes how to integrate this data into a research report.

The proposed approach is driven by research questions and has three stages, each related to the source of information, which is 1) The Internet, 2) Social Network Sites, and 3) Online Community. The Figure 1 contains a sketch of the entire approach, showing the researcher driven by his research questions and using the appropriate data sources (Literature and The Internet) to consolidate the findings in a scientific report. As a starting point, a researcher can go to the academic literature, establishing a base with which to begin his research quest. Notably, this approach leads to studies characterized by broad exploratory research in the early stages, which is incrementally refined into a single online community analysis. At all research stages, the researcher can have insights, which may affect or even change the original research questions, a characteristic of data-driven research. Finally, the literature review is also important to anchor the research findings in established theories in order to contribute to the existing body of scientific knowledge.

In the first stage, the researcher should utilize the facilities of the big search engines. Such search engines provide access to systems that provide information about how people are searching on the Web. Google Insights for Search[3] allows one to compare search volume patterns across specific regions, categories, time frames and properties. Yahoo! Search Clues[4] provides an exploratory environment to find interesting patterns in what people are looking for on Yahoo! Search. Similarly, Bing formerly provided access to Bing Trends[5], but Microsoft recently shut down the service. Other services can also be used to acquire mass data information, such as those cited by Eysenbach [Eysenbach 2009] (e.g. Google Ads).

The second stage happens at the level of social network sites, which are systems like Facebook, Twitter and Orkut. A researcher can use keywords identified in the previous stage to investigate what people are saying about the research topic. The results can provide an index of user interest toward the research themes (i.e. trends). The results

[3] http://www.google.com/insights/search

[4] http://clues.yahoo.com

[5] http://www.bing.com/trends

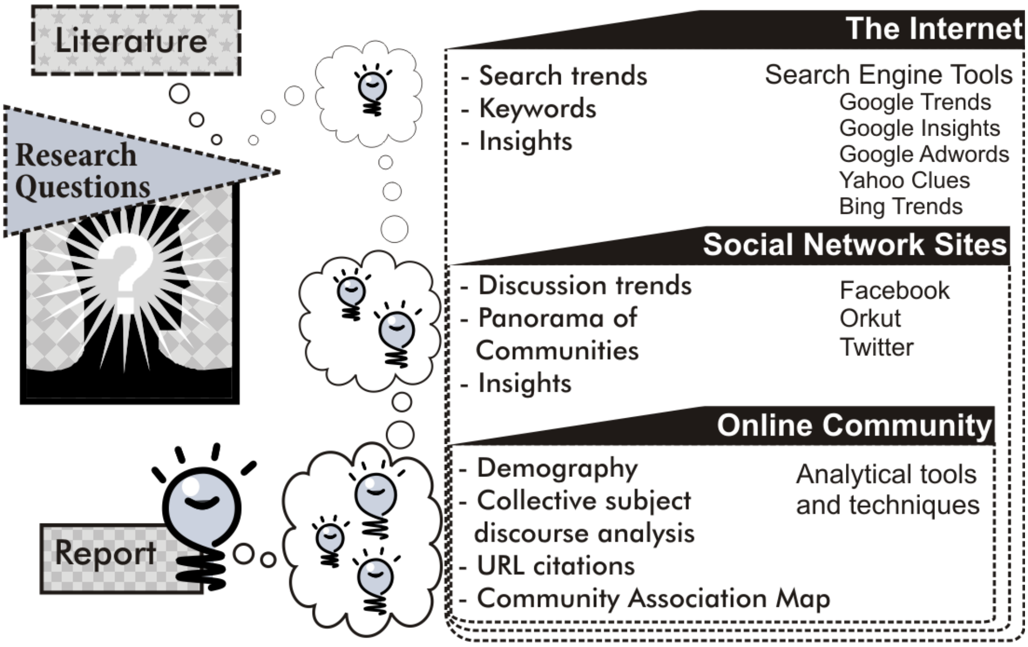


Figure 1. The proposed research approach

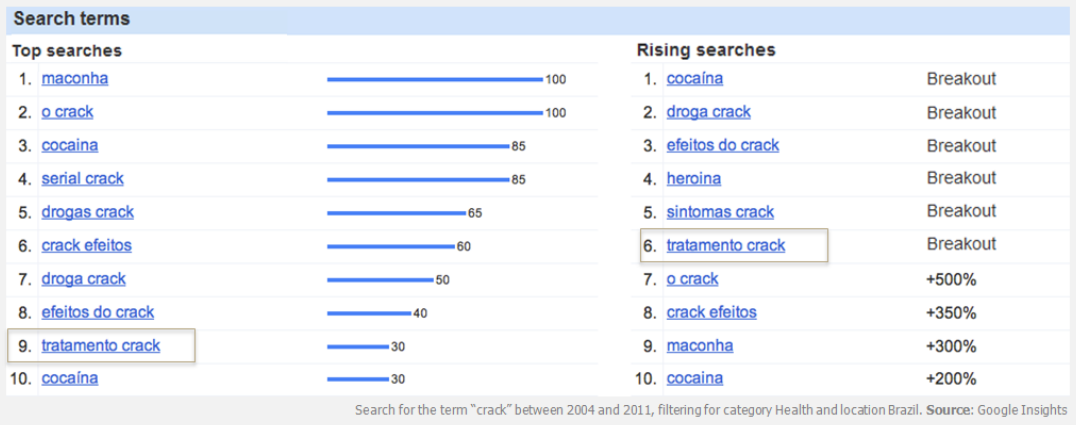
should also be factored into the decision making process about which community to study in depth. Tailored tools should be utilized in actual data processing and analysis. The next section shows an example of metrics to be used for this purpose.

The third and final stage takes an in-depth look at an online community, searching for answers to the research questions. As a way to understand the community, the researcher can perform a content analysis of discussion available at the community forum. A technique that aids in performing this analysis is the "Discourse of the Collective Subject", a qualitative technique with roots in the Theory of Social Representations [Lefevre and Lefevre 2006]. The aim of this technique is to identify collectives, aggregated by central ideas, and describe them through a discourse generated from a patchwork of members' speech, synthesizing the discourse as one collective subject. Another example of an analytical tool is the Community Association Map [Carvalho et al. 2012], which aims to help researches to analyze the users' interest in other communities. This tool can reveal patterns in interest and reinforce conjectures about the users' beliefs. Other tools can be used applied to help extract information from the community, for instance, URL extractors, and demographic data descriptors.

4. Example

As an application of the proposed approach, this section presents some results from a study about motivations for drug abuse to start and cease, specifically with regard to the drug crack cocaine in Brazil. The initial research questions were: 1) why do people start to abuse drugs; 2) why do they continue abusing them; and 3) why do they cease to abuse them. All three stages were conducted, providing a descriptive panorama about drugs on the Internet, especially with regard to the Brazilian Internet audience, and revealing the reality of a support community of users of crack cocaine. As a result of the community content analysis, the report compiled answers to the follow questions: 1) what are the leading factors to crack; 2) what are the optimal turning points at which to start a treatment; 3) what are abstinence maintenance factors; 4) what favors the drug abuse to restart; 5) what criticism exists for official health treatment; and 6) which kind of help are the codependents looking for.

Following the first stage recommendation, Google Insights for Search was used to get an overview of the search trends about the term "crack" in Brazil. Because of the ambiguity of the term "crack", which is also used by users looking for illegal software and licenses on the Internet, the configuration service was set to retrieve only searches related to the category of health. The Figure 2 shows the result of this query. Most of the searches are related to the general term crack, some searches are about the reaction to the drug ("crack efeitos"), and, as pointed out in this figure, many users are looking for information about treatment on the Internet ("tratamento crack") too.



Search terms

Top searches			Rising searches	
1.	maconha	100	1. cocaína	Breakout
2.	o crack	100	2. droga crack	Breakout
3.	cocaina	85	3. efeitos do crack	Breakout
4.	serial crack	85	4. heroina	Breakout
5.	drogas crack	65	5. sintomas crack	Breakout
6.	crack efeitos	60	6. tratamento crack	Breakout
7.	droga crack	50	7. o crack	+500%
8.	efeitos do crack	40	8. crack efeitos	+350%
9.	tratamento crack	30	9. maconha	+300%
10.	cocaína	30	10. cocaina	+200%

Search for the term "crack" between 2004 and 2011, filtering for category Health and location Brazil. **Source:** Google Insights

Figure 2. Google Insights for Search Results

Since the major source of social media in Brazil is Google's Orkut, this social networking site was used as the investigation platform for the stage two. A search for the term "crack" in the Orkut system, considering filters for location (Brazil) and language (Portuguese), gave 995 results in September of 2011. The next step was to select communities where content would have contextualized discourses about people's experience related to drug use. Categorizing the 995 communities, it was possible to identify 278 (28%) communities related to selection objective, 360 (36%) communities that seemed not to be directly related to the selection objective, and 357 (36%) communities that were not using the term "crack" in reference to a drug (e.g. instead referred to programs and password cracking). Narrowing the research, the 278 identified online communities were filtered to 13 communities considering other criterias such as: 1) possessing more than 300 members; 2) having been in existence for more than 6 months; 3) having exhibited recent activity; and 4) having content publically available. The last step of this stage was to choose one community for evaluation. The community "Crack, Nem Pensar - AJUDA"[6] was selected for in-depth analysis because, out of the 13 remaining communities, it is the oldest and has the most members (11,102). In a quick evaluation of its content, the community presented an intense conversation among its members, which in later analysis

[6] http://www.orkut.com/Main#Community?cmm=175318

showed an average of 3.3 messages per day since its creation in July 16 of 2004.

The community analysis focused on participating members who had engaged in conversation in the community forum. It is important to make this distinction, as all members have the potential to follow the discussions, but most choose not participate (i.e. lurkers). This analysis is based on the content of the participants who have posted messages in the forum. From the participant data available, there were 57% men and 43% women identified. The retrieved location of the participants was consolidated in a map (available online at `http://batchgeo.com/map/536db2e5aac00f746005efc6334542c4`). At the time of the study, September 2011, the community forum had 434 participants, 384 topics and 8655 messages, representing a total of 76.646 words, or 4.515.087[7] characters. The content analysis was conducted by applying the Discourse of Collect Analysis technique. Considering the high volume of data and efforts required for content analysis, a data cut was performed to focus the investigation in a suitable content analysis to the study objectives. From the original dataset, 39 (10%) topics were selected, with 129 (30%) participants and 925 (11%) messages, totalizing 107.488 (14%) words, or 602.332 (13%) characters.

The complete study results are currently being compiled in a journal article, and have been the subject of discussion in a seminar comprised of attendees from Brazilian government, health organizations and civil society. The community analysis identified that the speeches of dependents and codependents (the family and friends of dependents) are intermingled and complement each other, therefore both require care and attention. The reality of these people (i.e. life experiences recorded in the discussion) is transcript through discourse syntheses that answers to the study research questions. The purpose of the example was to give a glimpse of a holistic application of the approach, showing some illustrations and figures of what were found during the research.

5. Discussion

The proposed approach presents a practical guide to conducting studies in health base on data retrieved from the Internet. The main information source is based on online communities, but the search also begins with a broad view of the data available on the Web. Researchers should take care using these data, mainly because of questionable effects on the validity of the research findings. The study of online communities requires suitable content selection, based on information available at the source and other characteristics of these communities that are also available on the Web. Another important concern regards the ethical aspects of these studies. For that reason, this approach looks at public data available on the Web, suggesting the use of techniques that report aggregated data, therefore preserving the privacy of the individual user. In the literature, an extensive debate was found about these issues [Whitehead 2007, Kozinets 2009, Eysenbach 2009]. Although, several systems, services and tools support the execution of the approach, the proposed approach should be used as a guide for the realization of a study. Researchers may face dead-ends in data analysis due to size and time constraints; therefore, they should be creative and develop different ways to accomplish the analysis. It is best to consider the proposed approach as a practical road map to guide the data scientists through the research.

[7] The B-42 Gutenberg's Bible has around 3 million characters.

6. Conclusion

This paper introduced a practical approach by which to use public data available on the Internet for studies about healthcare issues. The proposed approach is an incremental three-step-in-depth process retrieving and analyzing information from the Web at different levels. The idea is to start analyzing data from search engine services that summarizes user search trends and information available on the Web. The next step utilizes social networking sites, looking for discussions related to the subject of interest based on keywords identified in the previous step. The final step provides a deeper analysis of a specific online community related to the research questions. The results presented in the example demonstrate the feasibility of the approach application based on its use in a real world study about drug use.

The application of communication and information technologies to health is a promising research theme. In this context, the development of new approaches and understandings of the ways in which people join forces, share and identify relevant information can foster improvement in our health quality. The proposed approach provides a guide for researchers conducting studies. Although available data and tools exist to support the execution of the approach in most studies, better tools and data can improve its execution. The evolution of research requires initiatives of opportunistic and inventive use of data and technology. The proposed approach is such an initiative, and more research is needed to establish better practices to aid in real problem solving.

ACKNOWLEDGEMENTS

The authors would like to thank Rodrigo Pazzini and Sérgio Rosa for the help with their expert social media skills and Elder Cirilo, Neven Ukrainczyk, Alorah Harman for their valuable discussion regarding this paper. This work has been sponsored by CNPq (Brazilian Council for Research and Development) – Bolsa de Doutorado CNPq 142620/2009-2.

References

Banks, A. (2011). The Brazilian Online Audience. Feb, 2011. URL: `http://www.comscore.com/por/Press_Events/Presentations_Whitepapers/2011/State_of_the_Internet_in_Brazil`. Accessed: 2012-04-01.

Bender, J. L., Jimenez-Marroquin, M.-C., and Jadad, A. R. (2011). Seeking support on facebook: A content analysis of breast cancer groups. *Journal of Medical Internet Research*, 13(1):e16.

Berger, M., Wagner, T. H., and Baker, L. C. (2005). Internet use and stigmatized illness. *Social Science & Medicine*, 61(8):1821–1827.

Boyd, D. M. and Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 13(1):article 11.

Carneiro, H. A. and Mylonakis, E. (2009). Google trends: A web-based tool for real-time surveillance of disease outbreaks. *Clinical Infectious Diseases*, 49(10):1557–1564.

Carvalho, D., Fuks, H., and Lucena, C. (2012). Community Association Map: Processing inter-community relationships. In *Proceedings of the 8th International Conference on Web Information Systems and Technologies (WEBIST)*, pages 665–670. Springer.

CGI (2011). *Survey on the use of information and communication technologies in Brazil: ICT Households and ICT Enterprises 2010*. Comitê Gestor da Internet no Brasil.

Cohen, R. A. and Stussman, B. (2010). *Health information technology use among men and women aged 18-64: Early release of estimates from the National Health Interview Survey, January-June 2009*. Health E-Stats. National Center for Health Statistics.

Eysenbach, G. (2009). Infodemiology and infoveillance: Framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the internet. *Journal of Medical Internet Research*, 11(1):e11.

Gold, J., Pedrana, A. E., Sacks-Davis, R., Hellard, M. E., Chang, S., Howard, S., Keogh, L., and andMark A Stoove1, J. S. H. (2011). A systematic examination of the use of online social networking sites for sexual health promotion. *BMC Public Health*, 11(1):583.

Gomide, J., Veloso, A., Meira, W., Almeida, V., Benevenuto, F., Ferraz, F., and Teixeira, M. (2011). Dengue surveillance based on a computational model of spatio-temporal locality of twitter. In *Proceedings of the ACM WebSci'11*, pages 1–8. ACM.

Greene, J. A., Choudhry, N. K., Kilabuk, E., and Shrank, W. H. (2011). Online social networking by patients with diabetes: a qualitative evaluation of communication with facebook. *Journal of General Internal Medicine*, 26(3):287–292.

IBGE (2011). *Sinopse do Censo Demográfico 2010*. Instituto Brasileiro de Geografia e Estatística.

Kozinets, R. V. (2009). *Netnography: Doing Ethnographic Research Online*. Sage Publications Ltd.

Lasker, J. N., Sogolow, E. D., and Sharim, R. R. (2005). The role of an online community for people with a rare disease: Content analysis of messages posted on a primary biliary cirrhosis mailinglist. *Journal of Medical Internet Research*, 7(1):e10.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., and Van Alstyne, M. (2009). Computational Social Science. *Science*, 323(5915):721–723.

Lefevre, F. and Lefevre, A. M. C. (2006). The collective subject that speaks. *Interface - Comunicação, Saúde, Educação*, 10(20):517–524.

Madeira, W. (2011). *Transformar é preciso: transformações na relação de poder estabelecida entre médico e paciente (um estudo em comunidades virtuais)*. PhD thesis, Faculdade de Saúde Pública, Universidade de São Paulo, BR.

Preece, J. and Maloney-Krichmar, D. (2005). Online communities: Design, theory, and practice. *Journal of Computer-Mediated Communication*, 10(4):article 1.

Telegraph, T. (2010). Archived by WebCite® at `http://www.webcitation.org/66b3jO4Fc`.

Whitehead, L. C. (2007). Methodological and ethical issues in internet-mediated research in the field of health: an integrated review of the literature. *Social Science & Medicine*, 57(4):782–791.