IDIOLECT-BASED IDENTITY DISCLOSURE AND AUTHORSHIP ATTRIBUTION IN WEB-BASED SOCIAL SPACES

Natalie Ardet
Freie Universität Berlin
Takustr.9, 14195 Berlin, Germany

Keywords: identity, social networks, privacy, authorship attribution, hypermedia, text mining.

Abstract: In this paper, we inspect new possible methods of Web surveillance combining web mining with sociolinguistic and semiotic related knowledge of human discourse. We first give an overview of telecommunication surveillance methods and systems, with focus on the Internet, and we describe the legal issues involved in Web or Internet communications investigations. We put the emphasis on identity disclosure and anonymity or pseudonymity undermining in open web spaces. Further, we give an overview of new trends in Internet mediated communication, and examine the virtual social networks they create. Finally, we present the results of a new method using the semiotic features of web documents for authorship attribution and identity disclosure.

1 INTRODUCTION

The Web is a growing social space allowing human interaction and communication. Thus legal and illegal telecommunication surveillance and user profiling methods have been developed since the emergence of the Internet and flourish on the Web. This raised the concern about privacy and anonymity in the field of computer mediated communication. We will describe some current aspects of Internet surveillance and depict the trends of communication in the Web, especially those communications taking place in *Web social spaces*. Finally, we will examine the problem of identity disclosure and anonymity (or pseudonymity) undermining in open web spaces. We illustrate the problem by presenting the results of a new approach using the properties of web documents to perform authorship attribution and thus leading to partial or complete identity disclosure.

2 TRADITIONS IN TELECOMMUNICATIONS SURVEILLANCE

In 1998, a study for the European Parliament (STOA, 1998) uncovered ECHELON, a multinational surveillance network, centered at Sugar Grove, WV, U.S.A, which intercepts all forms of electronic communications. It is the current electronic surveillance system used by the National Security Agency of the United States (NSA). This global system of electronic eavesdropping can monitor any electronic communication in the world. It was thought to be used for commercial purposes as well as military spying. ECHELON monitors millions of communications, uses computerized systems to target those of interest, by origin, destination, language or keywords. This electronic spy system was developed during the cold war. Targets are tagged and forwarded to Fort Meade for analysis and action. With the success of Internet connected personal computers, surveillance software targeting Internet users started to spread. *Spyware* consists of ”computer software that gathers and reports information about a computer user without the user’s knowledge or consent”(Wikipedia, 2005). In 2004, according to a study by the National Cyber-Security Alliance, 80% of home PCs are infested with spyware (AOL/NCSA, 2004). As such, spyware is cause for public concern about privacy on the Internet. The concern about *privacy* lead to a concern about the related concepts of identity, anonymity, pseudonymity, unlinkability and unobservability in computer mediated communication. Pfitzmann defines identity as ”any subset of attributes of an individual which uniquely characterizes this individual within any set of individuals” (Pfitzmann, 2004).

Ardet N. (2005).
IDIOLECT-BASED IDENTITY DISCLOSURE AND AUTHORSHIP ATTRIBUTION IN WEB-BASED SOCIAL SPACES.
In *Proceedings of the First International Conference on Web Information Systems and Technologies*, pages 305-310
DOI: 10.5220/0001234803050310

According to Jacobson, identity can be concealed by the mean of anonymity. While ”anonymity conceals an individuals real identity in online communication”, pseudonymity ”disguises” the real identity (Jacobson, 1999). An important factor which distinguishes pseudonymity from anonymity is the concept of *accountability*. Accountability is ”the property that ensures that the actions of an individual or an institution may be traced uniquely to that individual or institution” (ATIS, 2001). The major dilemma is that anonymity protects privacy but undermines accountability whereas identification assures accountability but threatens privacy (Clarke, 1999). Therefore Clarke stresses the ”central importance of pseudonymity as a primary means of achieving the necessary balance between the needs for privacy and for accountability” (Clarke, 1999). The *unlinkability* of a system is a property which ensures that a user of the system may make multiple uses of resources or services of the system without others being able to link these uses together (Pfitzmann and Köhntopp, 2001). In the following, the terminology in use is that of a setting in which ”senders send messages to recipients using a communication network”. We define the *unobservability* property as P_u and the anonymity property as P_a, s_i is a sender, r_j is a recipient, and R_{ij} is the communication relationship between s_i and r_j. According to Pfitzmann, we have following implications:

$$P_u \Rightarrow P_a \quad (1)$$
$$P_u(s_i) \Rightarrow P_a(s_i) \quad (2)$$
$$P_u(r_j) \Rightarrow P_a(r_j) \quad (3)$$
$$P_u(R_{ij}) \Rightarrow P_a(R_{ij}) \quad (4)$$
$$P_a(s_i) \Rightarrow P_a(R_{ij}) \quad (5)$$
$$P_a(r_j) \Rightarrow P_a(R_{ij}) \quad (6)$$
$$P_u(s_i) \Rightarrow P_u(R_{ij}) \quad (7)$$
$$P_u(r_j) \Rightarrow P_u(R_{ij}) \quad (8)$$

The concern of privacy led to new privacy-enhancing rules and technologies in order to refrain the exploitation and investigation of personal information in the Internet. An emerging example is the Platform for Privacy Preferences Project (P3P), a framework that allows users to control the amount of personal information they share with websites, developed by the World Wide Web Consortium (Cranor, 2002).

3 COMPUTER SUPPORTED SOCIAL NETWORKS

3.1 Social Networks and their analysis

What is a social network? According to Garton et al. (Garton, 1997), ”a social network is a set of people (or organizations or other social entities) connected by a set of social relationships, such as friendship, coworking or information exchange”. Still according to Garton et al. (Garton, 1997), a social network analysis is a structural analysis, which units of analysis is the interpersonal relation. A relation R (or *strand*) is characterized by its content c, direction d and strength s. A relation R denotes a relationship between two actors a and b. R is directed or undirected. An undirected relation can be unbalanced, which means that its expression is asymmetrical. A set of relations connecting a pair of actors is called a *tie*. The more relations (or strands) in a tie, the more *multiplex* (or multistranded) is the tie. The composition of a relation or a tie is derived from the social attributes of both actors. Wasserman and Faust (Wasserman and Faust, 1994) introduced a different definition of a tie, from their point of view a tie denotes a single aspect of a relationship. In the example in figure 1, the entity ”Web users” represents the actors. Each actor has a set of attributes, e.g. ethnicity, social class and gender. A possible relation between two actors is defined as ”interacts with”. This relation may have different aspects such as ”topic” or ”purpose of conversation”.

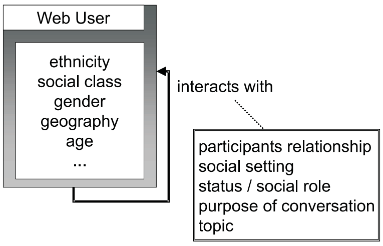

Figure 1: User interaction

3.2 Computer mediated communication

In the following part, we describe typical Internet computer mediated communication (CMC) and pinpoint those which create observable social networks. Email and Instant messaging systems (AIM, MSN,...) are basically point-to-point two-way communication, unless used for spamming purpose, then they are neither point-to-point nor two-way or communication. These means of communication create only transient social networks. We don't consider file-sharing systems because they don't support human to human communication based on naturally occurring text or speech. There are different types of one-to-many means of communication. Usenet is a decentralized system of discussion groups originally implemented in 1979-1980 by Steve Bellovin, Jim Ellis, Tom Truscott, and Steve Daniel at Duke University (Hauben and Hauben, 1997). Usenet predates the Internet, although today most Usenet material is distributed over the Internet using the NNTP protocol (Kantor and Lapsley, 1986). Estimated half of the existing Usenet newsgroups can be found on the Internet. Forums (or discussion boards) are centralized, usually web-based, discussion groups, allowing web users to have a ”conversation” about a given topic. A new kind of web communication support is the blog. Typically, a blog contains a chronologically ordered set of messages posted by a single author or a group of authors. Each message may be commented by its readers. The readers can link the blog to their own *blogroll* and thus creating a *blogspace*. The blogs are part of a web space called *blogosphere* (or blogsphere). A special feature of blogs and blogrolls is to export their content by the means of *feeds*. This technique is called content syndication. The most widely used syndication formats are currently *Really Simple Syndication* (RSS) and Atom (Nottingham, 2003). There are currently seven different RSS formats (Pilgrim, 2002). Blog aggregators like *Blogsnow*[1] syndicate the content of blogs. The aggregators are updated by a ping mechanism informing them about that an observed blog has been updated. The expansion of the weblogs has been documented by the blog survey company Technorati (see figure 2). Preece defines an online community as people, who interact socially, of a shared purpose, policies and computer systems to support and mediate social interaction (Preece, 2000). Internet researcher have created different taxonomies to describe those interactions. The terms *synchronous* is used for communication taking place in ”real time” and *asynchronous* for timely non continuous communication. Regarding online investigation, the emphasis is put on the transience or persistence of communication. Communication is persistent if it is archived in some way and can be recalled later. As an example, an email message represents a persistent communication unless you delete it. Communication need to be persistent and observable, to allow investigation. This is typical for Blogs and Usenet. Social spaces like discussion boards often reduce their observability by restricting their access and typically reduce the persistence of communication by the means of *pruning* techniques as showed in the virtual ethnography done by Ardet (Ardet, 2004).

[1] http://www.blogsnow.com

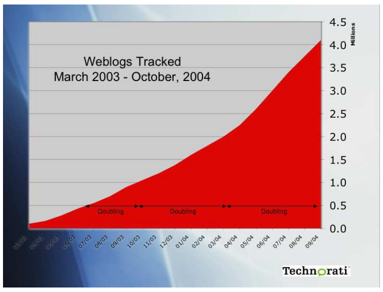

Figure 2: Weblogs usage

4 IDIOLECT AND SEMIOTIC FINGERPRINT

4.1 Properties of idiolects

Sociolinguistics is the study of language in human society. An aspect of sociolinguistic research is an area generally referred to as language variation. Language variation focuses on how language varies in different contexts, where context refers to factors like ethnicity, social class, sex, geography and age. The choice of the utterance used to convey a message is a also social decision tied to the social factors which define the relationship between the transmitter and the receiver.

Language, dialect and idiolect are additional factors which shape the utterances of the transmitter. A *dialect* is a collection of attributes that make one group of speakers noticeably different from another group of speakers of the same language. The term *accent* refers to a phonological variation which may be important in speech analysis. Accent is about pronunciation, while dialect is a broader term encompassing syntactic, morphological, and semantic properties as well. An *idiolect* is the variety of language spoken by each individual speaker of the language. While language and dialects are spoken by a group of person, an idiolect is only spoken by a single person. Chandler (Chandler, 2002) describes an idiolect as a term from sociolinguistics referring to the distinctive ways in which language is used by individuals. In semiotic terms it can refer more broadly to the stylistic and personal subcodes of individuals. As a result of these definitions, we can infer the following pyramid: a language consists of one or more dialects, and each dialect is a collection of idiolects as illustrated in figure 3. Each idiolect maps to only one person, which implies that there is an injective mapping function from an idiolect i to a person p. The mapping function cannot be a bijection because one person may speak different dialects from one or more language, depending if the person is monolingual or multilingual. We define X as a set of all idiolects and Y the set of all speakers. As f is injective, for every x and x' in X, whenever f(x) = f(x'), we must have x = x'. Which means it is possible to identify a person by the mean of his or her idiolect.

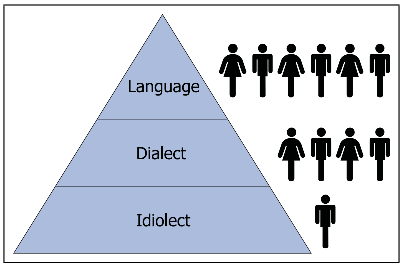

Figure 3: Language variety

4.2 Semiotic fingerprint

We now introduce the notion of a semiotic fingerprint, a content dependent property of a document. Considering $\alpha_i \in \mathbf{A}$ ($1 \leq i \leq n$) as a semiotic sign, we define the semiotic fingerprint function f_s as a mapping between a document $d = (\alpha_1, \alpha_2, \ldots, \alpha_n)$ and its semiotic fingerprint $\sigma_d = (\sigma_{d,1}, \sigma_{d,2}, \ldots, \sigma_{d,m})$ and $\sigma_{d,i} \in \mathbf{R}$, $\sigma_d \in \mathbf{F} = \mathbf{R^m}$:

$$f_s : \mathbf{D}^n \rightarrow \mathbf{F}$$
$$f_s(d) = \sigma_d$$

σ_d is a feature set of d, i.e. a vector representation of weighted and normalized attributes of d. Concerning the input of the semiotic fingerprint function f_s, i.e. the lexical unit of analysis of the forthcoming algorithm, it may be a single document or a set of documents. If the semiotic fingerprint function f_s is injective, i.e. $f(v_{t1}) = f(v_{t2})$ implies $v_{t1} = v_{t2}$, this would mean that each document has its own semiotic fingerprint, which means a semiotic fingerprint uniquely identifies a text. However, our goal is to define a semiotic fingerprint function which co-domain can be split in semiotic fingerprint classes $C_a \in \mathbf{F}$. Each class mapping eventually to an author:

$C_a = (\sigma_d \in \mathbf{F} | \exists d, f_s(d) = \sigma_d \wedge a = author(d))$

5 BLOGMINER WORKBENCH

5.1 Design and Conception

The BlogMiner workbench was designed to explore weblogs, a category of open social spaces in the Web. The discourse occurring in a weblog is encapsulated in messages posted on the blog. The publishing of messages usually occurs through the weblog's content management system (CMS). BLOGMINER fetches the discourse occurring in a selected weblog via the weblogs'own syndication mechanism. An algorithm to calculate the semiotic fingerprint of a message as defined in 4.2 has been developed and integrated in BLOGMINER. The semiotic fingerprint computing will be presented here. At first we have to define the set of features, which will be used in our analysis. The semiotic fingerprint data structure (SFDS) consists of two parts. The metrical part contains the vector containing the values for the semiotic fingerprints metrics (SFM). The lexical part contains the occurrence set of semiotic units, i.e. the distributional pattern of the words occurring in the unit of analysis. The metrical part may consists of occurrences and frequency of lexical, grammatical, or paralinguistic patterns (Meyer, 2001)(Ha, 2003). From a linguistic point of view, we have chosen an empiric approach. This is due to the ”fuzziness” of language occurring in many context of CMC as described in (Ardet and Thome, 2004). Therefore, our semiotic fingerprint doesn't include grammatical patterns. The inclusion of grammatical pattern in order to distinguish writing styles has been shown previously (Baayen et al., 1996) (Stamatatos and Kokkinakis, 2001) where the encoding of syntactic information has been done with part-of-speech n-grams. The lexical patterns we use include the frequency of punctuation, numbers, capital letters. The paralinguistic patterns we use are part of following categories:

Environmental contrast P_{ec} refers to a contrast between an object and its surroundings, caused by a difference in shape, color, or illumination (PLAI, 2005)

Affect indicator P_{ai} refers to pattern which aim to compensate the lack of facial expression or gesture in CMC (Ardet and Thome, 2004)

Hypermedia usage P_{hu} refers to the usage of hypermedia elements such as embedded images or hyperlinks.

Statistical values about the corpora like *word count*, i.e. the size of the corpus, *word types*, i.e. the vocabulary of the corpus and *word tokens*, i.e. the frequency of each word type, are stored in the second part of the semiotic fingerprint, the lexical part. These metrics have been discussed previously and thus have not been integrated in our approach yet (Oakes, 1998)(Smith, 1983)(Stamatatos and Kokkinakis, 2001)(Holmes, 1994) .

5.2 Results

The results achieved with our method for idiolect recognition based on semiotic fingerprint matching will be presented now. The semiotic fingerprint has been computed for messages of three distincts

Table 1: Fingerprint values

	σ_1	σ_2	σ_3	σ_4	σ_5	σ_6	σ_7	σ_8	σ_9	σ_{10}	σ_{11}	σ_{12}	σ_{13}	σ_{14}	σ_{15}	σ_{16}
σ_{E3}	3.2	4.6	1.8	0	9.2	2.7	3.7	0	0.9	0.9	7.4	10.1	1.8	0.9	0	0
σ_{E2}	8.6	4.5	2.2	0	4.5	2.2	2.2	0	1.1	0	0	1.1	2.2	0	0	0
σ_{E1}	1.8	4.6	3.1	0	12.5	3.1	4.6	0	0	1.5	0	15.6	7.8	1.5	0	0
σ_{M3}	8.3	7.6	0.1	0.8	2.6	0.5	0.4	0.8	0.1	0.2	0.7	0.8	0.2	0.1	3.8	0.7
σ_{M2}	3.8	4.4	0.3	1.2	5.0	1.2	1.2	2.2	0.9	0.3	0	0.9	1.2	0	1.9	0.6
σ_{M1}	6.1	6.0	2.3	1.1	6.3	2.0	2.0	1.4	1.4	0.2	3.4	0.8	0	0	2.6	1.4

weblogs[2]. In figure 4, each row shows semiotic fingerprints of a distinctive weblog. The weblogs messages are quite small (less than 200 words). Previous research result, couldn't achieve satisfactory results with small documents (De Vel et al., 2001a)(De Vel et al., 2001b). In figure 4, we can easily visualize the difference between the fingerprints in row 2, corresponding to weblog σ_{B_i} and the other fingerprints σ_{E_i} and σ_{M_i}. The semiotic fingerprints of σ_{E_i} and σ_{M_i} are visually quite similar, thus if we inspect the semiotic fingerprint values in table 1, we observe that in σ_{E_i}, the attributes $\sigma_4, \sigma_8, \sigma_{15}, \sigma_{16}$ are null whereas these values are not null in σ_{M_i}. The preliminary results on very small documents are quite promising and need some further investigation.

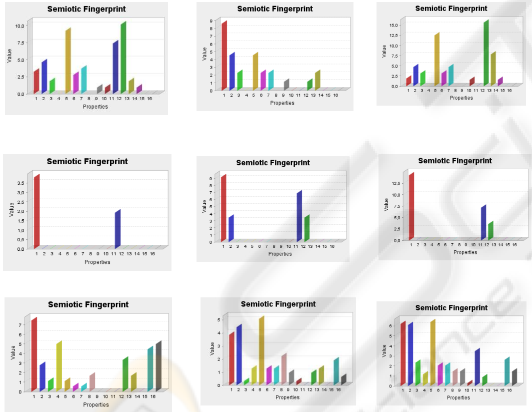

Figure 4: Example of Semiotic Fingerprints

6 RELATED WORK

Ongoing research in the field of authorship has led to some interesting results. Koppel et.al. showed that partial identity disclosure can be achieved using gender inference techniques (Koppel et al., 2003). Investigation of large corporas (with an average of 42000 words) have shown that several classes of simple lexical and syntactic features differ substantially according to author gender, especially regarding the use of pronouns and certain types of noun modifiers. Beside research concerning partial identity disclosure, the investigation of groups of interest, i.e. groups of individuals who share social spaces may also reveal information about an individual. In the context of *social network investigation*, the concept of group is useful for examining the relationships between sets of actors instead of single actors. Wellmann (Wellman, 1997) defines a group as "a social network whose ties are tightly-bounded within a delimited set and are densely-knit so that almost all network members are directly linked with each other". A particular group-finding algorithm is known as Friend-of-Friend (FoF) . This technique was first used in astrophysics by Huchra and Geller (Huchra and Geller, 1982) to identify group of galaxies. An entity belongs to a FoF group if it lies within some linking length ϵ of any other entity in the group. If no group is found, the entity is entered in a list of isolated elements. All entities found are added to the list of group members. The surroundings of each group member are then searched. This process is repeated until no further members can be found. Another approach to group-finding entitled *conversation maps* has been developed by Sacks. This approach aims to visualize the interaction between users of the Usenet in order to investigate very large conversations (Sack, 2000).

7 CONCLUSION

In this paper, we first explored the technical aspects of the Internet, considered as a social model, as suggested by the psychological view of Mantovani (Mantovani, 2001). Then, we gave a definition of computer supported social networks, described their topology (centered or distributed), their usages (personal and business oriented) and explored the topological properties of two emerging types of computer-supported social networks, aka Web social spaces. Finally, we illustrated the issues of privacy and identity disclosure in open Web social spaces by presenting the preliminary results of an algorithm for idiolect recognition

[2] blabbermouth.net, executivewoman.blogspirit.com and motorcitybadkitty.com

based on semiotic fingerprint matching. Our approach uses the rich media content of web documents to build a set of features, thus allowing stylometric analysis on small documents.

REFERENCES

AOL/NCSA (2004). Aol/ncsa online safety study. Technical report, America Online and the National Cyber Security Alliance.

Ardet, N. (2004). *Teenagers, Internet and Black Metal music*. Conference Proceedings CIM 2004.

Ardet, N. and Thome, M. (2004). *Virtual Ethnography: a Computer-Based Approach*. Computer and their Applications, Conference Proceedings (2004), 79–82.

ATIS (2001). Alliance for telecommunications industry solutions telecom glossary.

Baayen, H., Halteren, H., and Tweedie, F. (1996). Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing, 11*.

Chandler, D. (2002). *Semiotics: The Basics*. London, Routeledge.

Clarke, R. (1999). Identified, anonymous and pseudonymous transactions: The spectrum of choice. *User Identification & Privacy Protection Conference, Stockholm*.

Cranor, L. F. (2002). *Web Privacy with P3P*. O'Reilly & Associates.

De Vel, O., Anderson, A., and Corney, M. (2001a). Mining e-mail content for author identification forensics. *ACM Sigmod, Volume 30 , Issue 4 (December 2001)*.

De Vel, O., Andersond, A., and Corney, M. (2001b). Multi-topic-e-mail authorship attribution forensics. In *ACM Conference on Computer Security - Workshop on Data Mining for Security Applications, November 8, 2001, Philadelphia, PA, USA.*

Garton, L. (1997). Studying on-line social networks. *JCMC (Journal of Computer Mediated Communication) Vol.3, Issue 1, 1997*.

Ha, L. A. (2003). Extracting important domain-specific concepts and relations from a glossary. In *Proceedings of the 6th CLUK Colloquium*, pages 49–56, Edinburgh, UK.

Hauben, M. and Hauben, R. (1997). *Netizens: On the History and Impact of Usenet and the Internet*. Wiley-IEEE Computer Society Press.

Holmes, D. I. (1994). Authorship attribution. *Computers and the Humanities*, Nr. 28:87–106.

Huchra, J. and Geller, M. (1982). *Groups of Galaxies I. Nearby Groups*. ApJ 257 423.

Jacobson, D. (1999). Doing research in cyberspace. *Fields Methods*, Vol. 11, No. 2, November 1999:pp. 127–145.

Kantor, B. and Lapsley, P. (1986). Network news transfer protocol. Technical report, U.C. San Diego.

Koppel, M., Argamon, S., and Shimoni, A. (2003). *Automatically categorizing written texts by author gender*. Literary and Linguistic Computing 17(4), November 2002, pp. 401–412.

Mantovani, G. (2001). The psychological construction of the internet. from information foraging to social gathering to cultural mediation. *Cyberpsychology And Behavior. Vol. 4 (1), Pp. 47-56.*

Meyer (2001). Extracting knowledge-rich contexts for terminography. In D. Bourigault, C. J. and LHomme, M. C., editors, *Recent Advances in Computational Terminology*. Amsterdam, John Benjamins.

Nottingham, M. (2003). The atom syndication format 0.3 (pre-draft). Technical report, Atom Working Group.

Oakes, M. P. (1998). *Statistics for Corpus Linguistics*. Edinburgh.

Pfitzmann, A. (2004). Anonymity, unobservability, pseudonymity, and identity management a proposal for terminology (draft v0.21 sep. 03, 2004). Technical report, TU Dresden.

Pfitzmann, A. and Köhntopp, M. (2001). Anonymity, unobservability, and pseudonymity a proposal for terminology. Technical report, proposal.

Pilgrim, M. (2002). What is rss? *www.xml.com*.

PLAI (2005). The plain language association international glossary. http://www.plainlanguagenetwork.org/.

Preece, J. (2000). *Online communities*. Wiley.

Sack, W. (2000). Conversation map: A content-based usenet newsgroup browser. In *in the Proceedings of the International Conference on Intelligent User Interfaces (New Orleans, LA: Association for Computing Machinery, January 2000)*.

Smith, M. (1983). *Recent Experience and New Developments of Methods for the Determination of Authorship*. Association for Literary and Linguistic Computing Bulletin, 11, 1983, S. 73-82.

Stamatatos, E., N. F. and Kokkinakis, G. (2001). Computer-based authorship attribution without lexical measures. *Computers and the Humanities 35*, pages 193–214.

STOA (1998). An appraisal of technologies of political control, interim study. Technical report, STOA Programme, Directorate-General for Research Directorate B, Eastman 112, rue Belliard 97-113, B-1047 Bruxelles., http://cryptome.org/stoa-atpc.htm.

Wasserman, S. and Faust, K. (1994). *Social Network Analysis: Methods and Applications*. Cambridge University Press.

Wellman, B. (1997). *Cultures of the Internet*, chapter An electronic group is virtually a social network, page pages 179205. Lawrence Erlbaum Publications, Mahwah, New Jersey.

Wikipedia (2005). *Wikipedia Encyclopedia*. www.wikipedia.com.