Mental Models of Data Privacy and Security Extracted from Interviews with Indians

Jana Diesner, Ponnurangam Kumaraguru, Kathleen M. Carley

Abstract

The Indian software and services market continues to gain momentum, with offshore outsourcing from the US, Europe and other countries becoming mainstream. As jobs that involve processing of personal data are increasingly outsourced to India, concerns are being raised about the protection of this data. While a large number of studies has been conducted in order to assess people's attitudes about data privacy and security in the US, Australia, Canada and Europe, little information is available on this topic in India.

The research we present seeks to gain an empiric and exploratory understanding of Indians' attitudes about data privacy and security. We study these attitudes by analyzing the mental models that are reflected in interviews which we conducted among Indians. We will report on a methodology for extracting, analyzing and comparing mental models from texts and on the knowledge we gained about the perception of data privacy and security among the subjects.

Keywords:

Mental Models, Network Text Analysis, Data Privacy, Data Security, India, AutoMap

Direct all correspondence to Jana Diesner, Institute for Software Research International, Carnegie Mellon University 5000 Forbes Avenue, Pittsburgh, PA, 15213; e-mail: diesner@cs.cmu.edu. This work was supported in part by the National Science Foundation IGERT grants 9972762 in CASOS, the CASOS (Center for Computational Analysis of Social and Organizational Systems) lab at CMU (http://www.casos.cs.cmu.edu), and the CyLab at CMU (http://www.cylab.cmu.edu). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the National Science Foundation or the U.S. government. We thank Lorrie Cranor for her support and advice on this research and publication. We thank Elaine Newton and Granger Morgan for creating the questionnaire, Pranjali Kanade for transcribing the interviews, and Dan Woods for his help with this paper.

Mental Models of Data Privacy and Security Extracted from Interviews with Indians

1. Introduction

The Indian software and services market continues to gain momentum, with offshore outsourcing from the US, Europe and other countries becoming mainstream. As jobs that involve processing of personal data are increasingly outsourced to India, concerns are being raised about the protection of this data. While a large number of studies has been conducted in order to assess people's attitudes about data privacy and security in the US, Australia, Canada and Europe [8][10][18][13], little information is available on this topic in India. India currently has no laws on data privacy, but new privacy laws are being discussed.

The research presented herein seeks to gain an empiric and exploratory understanding of Indians' attitudes about data privacy and security. We study these attitudes by extracting and analyzing mental models of the perception of data privacy and security from interviews conducted among Indians. We refer to mental models as representations of the reality that people have in their mind and use to make sense of their surroundings [11][15]. Herein we study mental models as cognitive constructs [12] that reflect the subjects' knowledge and information about data privacy and security. In this paper we report on the methodology we used for extracting, analyzing and comparing mental models from texts as well as on the knowledge we gained about the subjects' perception of data privacy and security. Our results might be of use for supporting the development of privacy-related policies and technologies that are tailored for the Indian environment.

2. Interview Protocol and Data

The interview protocol that we used aimed to gather information on attitudes of Indians towards a wide range of aspects of data privacy and security. The interviewees were asked 17 open questions on the following four topics: 1) A person's general understanding or mental models of data privacy and security. 2) Knowledge of privacy risks (e.g. data sharing and selling) and protections against such risks. 3) Awareness of Internet privacy and security issues. 4) Concerns about computerization of data.

Furthermore, we collected demographic information (excluding personal data that would allow for the re-identification of interviewees). We contacted a sample of potential subjects who had shown interest in participating in such interviews or discussions. Criteria for subject selection were an age of at least 23 years, possession of a Bachelor's or higher degree, and working experience of at least six months or current employment.

We interviewed 29 subjects (one at a time), recorded the conversations and transcribed the data to text files. The transcripts contain the interviewees' answers to the questions asked by the interviewer, but not the questions themselves. We note that the topics raised and the answers given are impacted by the questions that the interviewer asked. The resulting corpus contained 29 texts, a total of 57,092 words, and 3,729 unique words. The number of unique words considers each distinct word once per corpus, whereas the total number of words also considers repetitions of unique words. Table 1 provides basic statistics on the data set.

Interviews were conducted among subjects from different states of India. The subjects' ages ranged from 23 to 65; 59% of the individuals were between 26-35 years old. 62% of the interviewees were male, and 38% of the subjects held a Masters Degree. 62% of the participants were qualified in non-technical fields such as Linguistics, Arts or Accounting.

3. Methodology

We performed Network Text Analysis (NTA) on our corpus in order to extract the individuals' mental models from the texts. NTA is a set of techniques that is based on the insight that language and knowledge can be modelled as the network of words and the relations between them [16][14]. Extracting and analyzing networks of linked concepts is therefore a way to get to the meaning of texts [7]. In this study we use a specific NTA technique called map analysis, which systematically reveals the network of ties between words in a text in order to represent the author's "mental model" as a map [5][6][7]. In map analysis, a *concept* is an ideational kernel such as "data privacy" or "internet." A *statement* is two concepts and the relation between them. Statements are also referred to as links. A *map* is the network of all statements per text. The Map Analysis technique has been formalized and implemented into the AutoMap software [9]. We used AutoMap-2.0.12 for this study.

Before running NTA we pre-processed the data in AutoMap in order to condense the texts to the concepts that re most relevant for studying mental models of data privacy and security. We used the pre-processing technique of generalization, which translates text level concepts that represent relevant content into higher level concepts that represent the text level concepts in a generalized way [4]. The content of the thesaurus depends on the corpus and domain. We identified the higher level concepts by applying our previous knowledge about data privacy and security. Also we looked at the distribution of concepts across the corpus and selected relevant terms that were frequently used. Text level concepts include N-grams such as "credit card" or "social security number." The thesaurus was developed incrementally. This means that after each phase of extension and refinement we applied the thesaurus to the data and checked if further additions or modifications needed to be made in order to cover the terms that were relevant for studying data privacy and security. The resulting thesaurus contained 818 associations of unique text level concepts with 148 unique higher level concepts (see Appendix for the higher level concepts). The thesaurus creation required about 3 hours of manual work from 3 people. Table 1 provides statistics on the thesaurus and its impact on the data.

We applied the thesaurus in AutoMap in such a way that only higher level concepts that text level concepts had been translated into were maintained in the generalized texts. All other concepts were disregarded and replaced with imaginary placeholders that ensure the maintenance of the original distance of the translated terms [9]. Table 1 provides statistics on the outcome of this procedure.

Table 1: Statistics of translated texts

	Sum across corpus	Average per text	Minimum	Maximum	Standard Deviation
Unique concepts in original texts	13,919 (accumulated from 3,792 unique ones)	480	176	771	158.5
Total concepts in original texts	57,092	1898.2	355	3591	829.0
Higher level concepts in thesaurus	148	5.5	1	47	5.7
Unique concepts in translated texts	1,896 (accumulated from 148 unique ones)	65.4	38	104	15.6
Total concepts in translated texts	8,393	289.4	83	525	119.2

After pre-processing the data we specified the statement formation settings which determine how concepts are linked into statements (for detailed description of statement formation see [9]). In order to find the statement formation setting that best matched our data as well as the links that human coders identified we ran several pre-tests with various statement formation settings. After each pre-test we randomly picked an input text, had two independent human coders coding a portion of this text, and compared the hand coding results against the machine generated results. The human coders were using the same thesaurus used by AutoMap. The purpose of the strategy was to find the statement formation setting that most closely resembled human coding. Based on the insights we gained from the pre-tests we decided to form links within, but not across sentences, using a window size of eight. With this setting we covered the statements that hand coders were finding within a sentence. The window size is the maximum distance of concepts that will be linked into statements. It seemed that in this dataset interviewees used sentences (as supposed to an entire answer block or paragraph) as the unit in which they expressed a distinct idea.

After pre-processing the data and specifying the statement formation settings we ran Map Analysis on the entire corpus. We did not code the valence of a link; meaning that the resulting maps do not indicate if a statement had a positive or negative connotation for the interviewee.

4. Results

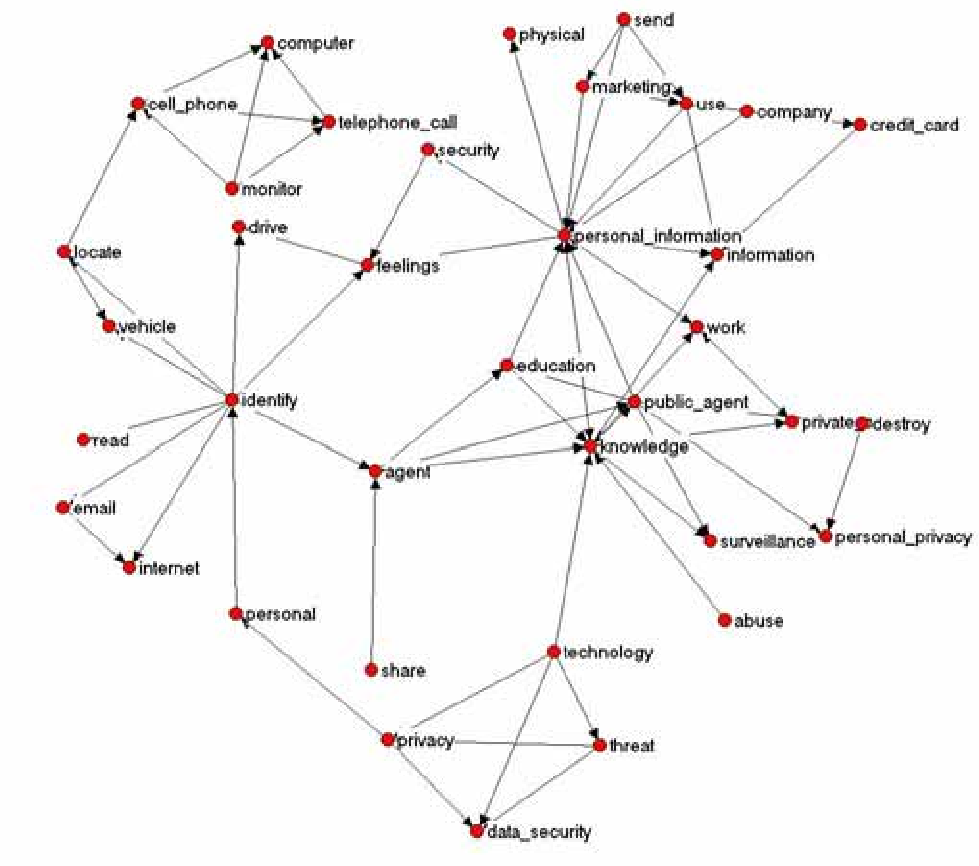
We extracted one map per text. The statements in the networks are weighted by the frequency of the occurrence of each statement per text. Table 2 provides basic statistics on the networks.

Table 2: Statistics of extracted networks

	# concepts analyzed unique	# concepts analyzed total	# concepts in statements unique	# concepts in statem. total	# isolates unique	# isolates total	# Statements unique	# Statements total	Density unique	Density total
Average	65.4	289.4	181.5	228.8	29.3	71.1	181.5	228.8	2.6	3.3
Min	38.0	83.0	42.0	51.0	14.0	29.0	42.0	51.0	1.1	1.3
Max	104.0	525.0	359.0	483.0	49.0	116.0	359.0	483.0	3.8	5.4
StdDev	15.6	119.2	80.7	107.6	9.4	20.2	80.7	107.6	0.7	1.0

We output the maps in the DL network representation format in order to load them into NetDraw [2] and visualized them. Figure 1 gives an example for the mental model from one text.

Figure 1: Mental Model from one text



Such a graph tells the analyst which concepts are connected to each other, the direction of the link, and which concepts do not form links, thus being isolates. From looking at all maps we learned that the maps have very few components. Disregarding isolates most maps have only one component (see Figure 1 for example). Components are maximally connected subset of nodes, also referred to as subgraphs [17]. In a component each node can be reached from any other node. The existence of components indicates that a graph is disconnected. Note that we are referring to weak components, meaning that the direction of a link is disregarded. The low number of components indicates that the subjects strongly interconnect the concepts that were discussed in the interviews. The maps also are very elaborate in terms of the number of concepts used and connected. Many portions within the components form cliques. Cliques are maximally complete subgraphs, meaning that each concept is connected to each other concept. Besides the cliques we also frequently observed a structure that we informally label as “hubs”: central concepts with terms grouped around them that may or may not be interconnected (for example “identity” and” personal information” in the graph above). Cliques can be interpreted as a set of ideas that are invoked concurrently, whereas hubs might serve as central reference points or starting points for a broader set of ideas that people relate to a central concept.

Next we created the union of all maps using CompareMap [9] in order to study the common set of ideas or the consensus represented in the subject’s perception of data privacy and security. Since the strength of statements per text has high variance (see Table 1), the cumulative sum of statements in the union file showed an even wider frequency distribution (ranging from 1 to 202). In order to normalize outlying texts with respect to the high frequency of certain statements we created binary maps in AutoMap and created the union of the binary maps. As a result, the strength of a link in the union file indicates the number of texts in which a statement occurs. The visualization of the network that contains each link that appears in at least one text results in a dense and hard to interpret ball of yarn. Considering links that occur in at least 20% (Figure 2), 33% (Figure 3), 50% (Figure 4), and 66% (Figure 5) of the maps provides a clearer picture:

Figure 2: Union of maps, links present in at least 20% of the maps

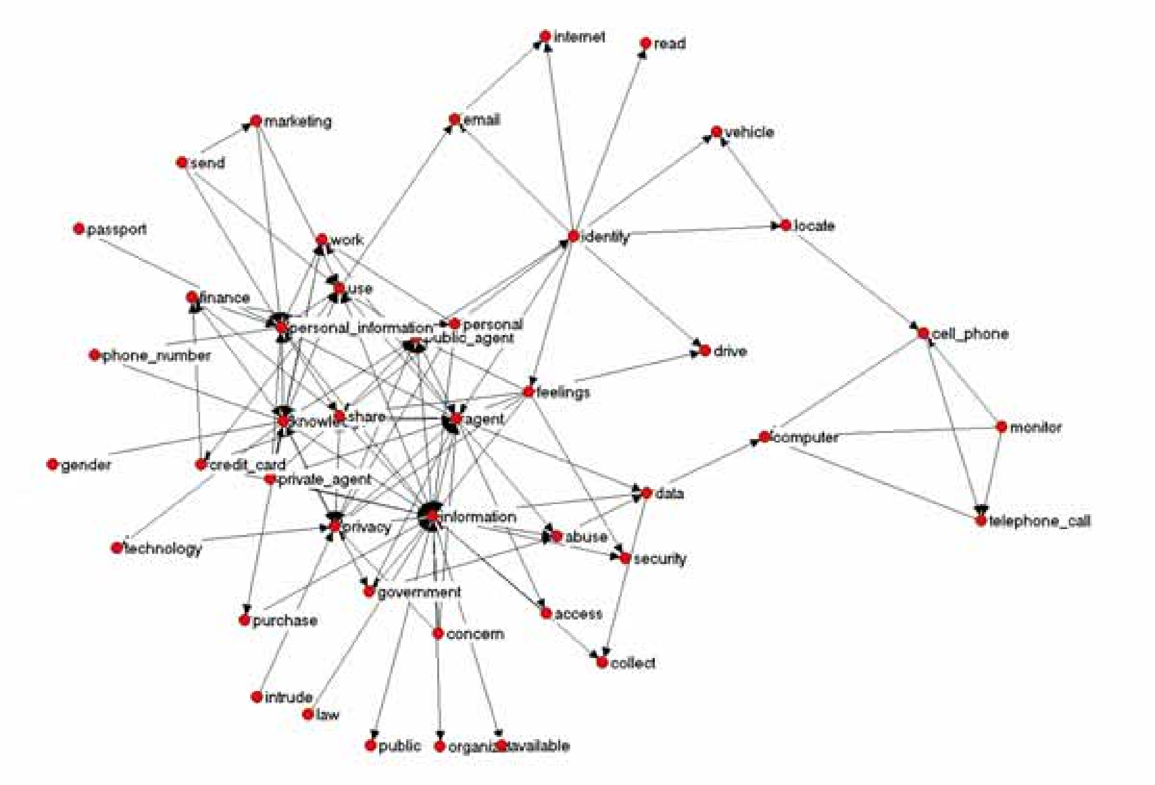
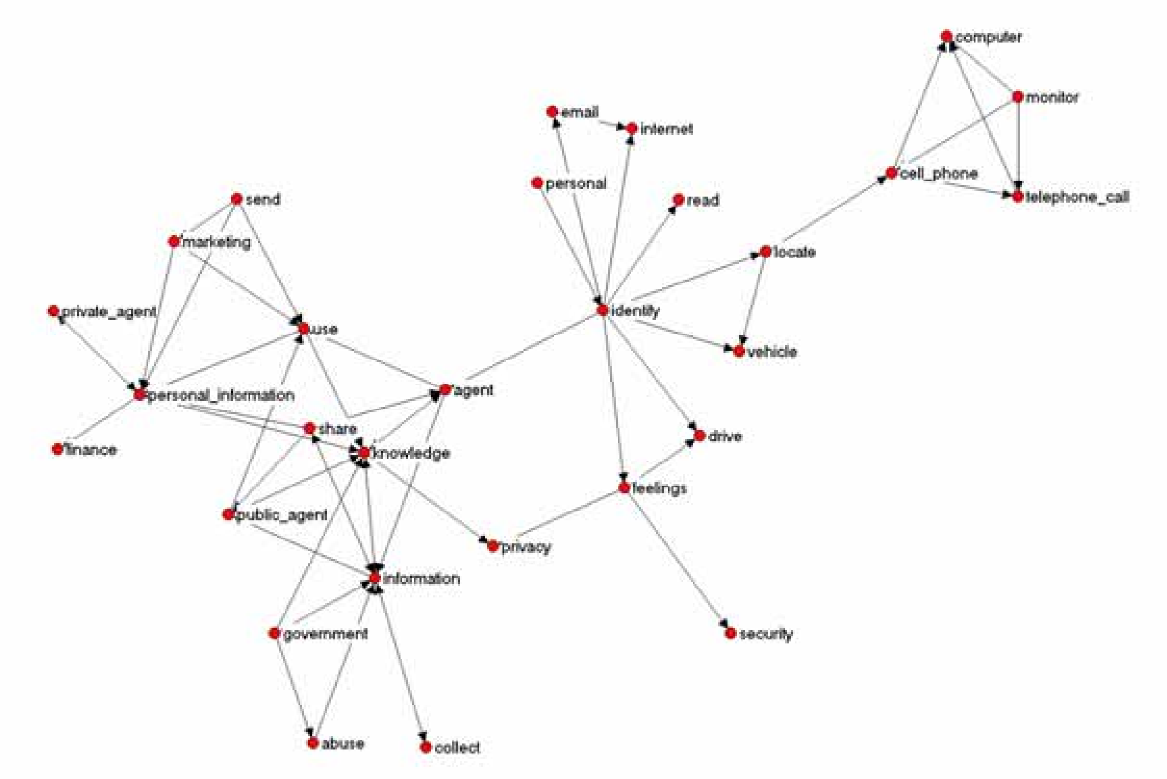
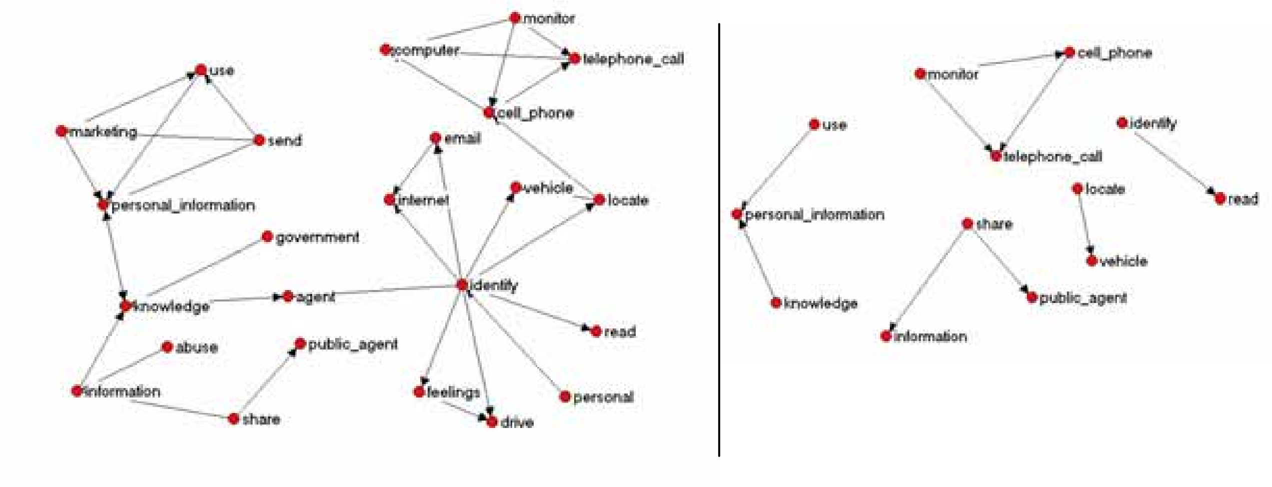


Figure 3: Union of maps, links present in at least 33% of the maps



Figures 4 and 5: Union of maps, links present in at least 50% and 66% of the maps



In those maps “personal information”, “identify” and “knowledge” appear as key concepts in the context of data privacy and security. Personal information is mainly related to the private sector such as “companies”, “marketing” and “finance”. The concept identify is more associated with personal features such as “email”, “internet”, “reading” and “feelings”. Statements that connect “cell phone” and “phone call” with “monitor” appear in a large proportion of the maps, indicating that people are aware of this possible security threat. Other links that indicate privacy and security threads however are not shared among many maps: Identical links that include “threat” are present in 5 maps, “identity theft” in 2, and “privacy concern” and in 1. “Abuse”, however, appears in 18 map and is related to information and government. The concepts of “personal information”, “identify”, and “knowledge” are much more elaborated than terms that imply a negative meaning. This finding suggests that the interviewees are not necessarily worried about data privacy and security issue or personal disadvantages that can result from the violation of data privacy and security.

The mental maps indicate that some concepts take more central positions in the networks than others. In order to understand which concepts are most central in people’s mental models of data privacy and security, we computed a set of centrality measures on the union of maps in UCINET [3] (Table 2). We argue that the most central concepts are the strongest representations of the interviewee’s idea of data privacy and security.

Table 3: Most central concepts

Centrality Measures				
Degree	**Closeness**	**Betweenness**	**Eigenvector**	**Overall Ranking**
Concept (Value)	**Concept (Value)**	**Concept (Value)**	**Concept (Value)**	**Rank Concept**
Agent (73.2)	Agent (44.1)	Agent (9.6)	Agent (28.9)	1. Agent
Information (69.0)	Information (43.3)	Information (8.8)	Information (27.9)	2. Information
Knowledge (66.9)	Knowledge (42.9)	Knowledge (6.6)	Knowledge (27.8)	3. Knowledge
Public Agent (59.9)	Public Agent (41.6)	Personal Info (4.8)	Public Agent (26.6)	4. Public Agent
Personal Info (57.7)	Personal Info (41.3)	Public Agent (4.4)	Personal Info (25.3)	5. Personal Info

Closeness centrality describes how close a concept is to all other concepts. Betweenness centrality measures how often a concept is positioned on the shortest path between any other pair of concepts. Eigenvector centrality tells the analyst how close a concept is to other concepts that are important with respect to degree centrality, and a degree is the number of other concepts directly linked to a concepts. The computation of centrality measures suggests a different set of more prominent concepts in the maps than the visual inspection of the maps did. From those results we learn that security related terms are not central concepts in people’s minds. The interviewees build their mental models around information, knowledge, people and personal information.

5. Discussion

The results shown in this paper are of an exploratory nature based on a small sample and therefore cannot be generalized. The absence of the valence of links restricts us from dividing ideas that are perceived in a positive way from negative connotations. However, the coding contains concepts that have a clear positive or negative meaning, which enables us to see how those terms are being used and related.

The software we used does not support the detection of negations. One would need a part-of-speech tagging based natural language processing package in order to take negations into account. We tried to circumvent this limitation by including negations into the thesaurus (e.g. don’t know, insecure).

6. Conclusion

The research presented herein demonstrates a methodology for extracting and analyzing mental models from texts via computer-supported NTA. We have shown that peoples’ mental models of a certain domain – in this case data privacy and security - can be studied at an individual level by looking at idiosyncrasies in a network, and on a group level by analyzing the properties and content of maps. We argue that the techniques used in this study can enable analysts to gain a quick, higher level understanding of the content of a dataset on various levels of aggregation. Furthermore the network analytic perspective and measure provide analysts with the ability to study not only what ideas are represented in a corpus, but also how people relate them to each other. The coding scheme can easily be adjusted as new data is being added to the corpus, and analyses with various coding schemes can be run quickly on the entire dataset.

From a content domain perspective we are providing a first empiric insight into the perception of data privacy and security among Indians. The mental models suggest that the interviewees center their perception of data privacy and security around the concepts of “personal information”, “identify”, and “knowledge”. “Cell phones” and “telephone calls” are being related to “monitor”, indicating a possible concern about this security threat. Concepts and statements that imply a negative attitude or concerns about data privacy and security in general however do not show as elaborate ego-maps as terms with a positive meaning.

In future work we plan to compute Hamming Distances between all maps in order to analyze if similar mental models correlate with similar demographic features of the subjects. Furthermore, we plan to compare our results with similar studies conducted in the USA. From such a comparison we hope to learn more about cross-cultural differences and similarities of current images of data privacy and security.

References

[1] Banks, D., & Carley, K.M. (1994). Metric Inference for Social Networks. *Journal of Classification*, 11, 121-149.

[2] Borgatti, S.P. (2002). *NetDraw1.0*.

[3] Borgatti, S.P., Everett, M.G., & Freeman, L.C. (2002). *UCINET for Windows.*

[4] Burkart, M. (1997). Thesaurus. In M. Buder, W. Rehfeld, T. Seeger, & D. Strauch (Eds.), *Grundlagen der praktischen Information und Dokumentation: ein Handbuch zur Einführung in die fachliche Informationsarbeit* (4th edition) (pp. 160 – 179). München, Germany: Saur.

[5] Carley, K.M. (1988). Formalizing the Social Expert's Knowledge. *Sociological Methods and Research*, 17(2), 165-232.

[6] Carley, K.M. (1997). Network Text Analysis: The Network Position of Concepts. In C.W. Roberts (Ed.), *Text analysis for the Social Sciences* (pp. 79-102). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

[7] Carley, K.M., & Palmquist, M. (1992). Extracting, Representing, and Analyzing Mental Models. *Social Forces*, 70(3), 601-636.

[8] Cranor, L., Reagle, J., & Ackerman, M.S. (1999, September 25-27). Beyond Concern: Understanding Net Users' Attitudes About Online Privacy. *Telecommunications Policy Research Conference*. Alexandria, VA.

[9] Diesner, J., & Carley, K.M. (2004). *AutoMap1.2 – Extract, analyze, represent, and compare mental models from texts*. Technical Report, Carnegie Mellon University, School of Computer Science, Institute for Software Research International, URL: http://reports-archive.adm.cs.cmu.edu/anon/ CMU-ISRI-04-100.pdf

[10] Harris, L., & Westin, A.F. (1998, June). *E-Commerce & Privacy: What Net users want*. Sponsored by Privacy & American Business and Price Waterhouse, Inc.

[11] Johnson-Laird, P. (1983). *Mental Models*. Cambridge, MA: Harvard University Press.

[12] Klimoski, R., & Mohammed, S. (1994). Team mental model: Construct or metaphor?. *Journal of Management*, 20, 403-437

[13] Morgan, R. (2001). *Privacy and the Community*. URL: http://www.privacy.gov.au/publications /rcommunity.html

[14] Popping, R. (2000). *Computer-assisted Text Analysis*. Thousand Oaks, London: Sage Publications

[15] Rouse, W. B., & Morris, N. M. (1986). On looking into the black box; prospects and limits in the search for mental models. *Psychological Bulletin*, 100, 349-363.

[16] Sowa, J.F. (1984). *Conceptual Structures: Information Processing in Mind and Machine*. Reading, MA.

[17] Wasserman, S., & Faust, K. (1994). *Social Network Analysis*. New York: Cambridge University Press.

[18] Westin, A., & Center for Social & Legal Research. (1970-2003). *Bibliography of surveys of the U.S. Public*. Retrieved August 15, 2004, URL: http://www.privacyexchange.org/iss/surveys /surveybibliography603.pdf

Appendix

Higher level concepts in generalization thesaurus:

abroad, abroad, abuse, access, accident, agent, available, avoid, aware, book, cell_phone, choice, collect, communication, company, computer, concern, connect, control, copy, copyright, credit_card, crime, data, data_privacy, data_security, decision, destroy, develop, development, disclaimer, disease, disturb, don't_know, download, drive, education, educational_agent, electronic_information, email, embarass, ethics, experience, exploit, extremist, feelings, filter, finance, firewall, foreign_government, freedom, future, gender, government, habit, harm, hear, hide, home, homosexual, identify, identity, identity_theft, India, Indian_government, information, insecure, insurance, intelligence, internet, intrude, irritate, knowledge, law, locate, market_agent, marketing, medical, mind, monitor, national_security, network, opinion, organization, organizational_agent, passport, patent, permission, personal, personal_information, personal_privacy, phone_number, photo, physical, police, policy, prevent, privacy, privacy_concern, privacy_problem, privacy_statement, private_agent, private_sector, problem, protect, protocol, public, public_agent, purchase, read, record, religion, remote, responsibility, restrict, right, security, send, share, shopping_habit, singapore, sms, social, social_security_number, societal_agent, societal_information, societal_privacy, society, sophisticated, speak, standard, store, surveillance, system, technical_privacy, technology, telephone_call, terrorism, terrorist, theft, threat, transaction, transparent, travel, trust, use, vehicle, walk, weapon, work