Big data, privacy, and trusted web: What needs to be done

Dirk Helbing and Stefano Balietti

ETH Zurich, ETH Zurich

30 May 2011

Big Data, Privacy, and Trusted Web: What Needs to Be Done[1]

Dirk Helbing[1,2] and Stefano Balietti[1]

1 ETH Zurich, CLU, Clausiusstr. 50, 8092 Zurich, Switzerland
2 Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA
∗ **E-mail:** dhelbing@ethz.ch and sbalietti@ethz.ch

Abstract

This perspective paper discusses challenges and risks of the information age, and the implications for the information and communication technologies that need to be built and operated. It addresses ethical and policy issues related with Big Data and how procedures for privacy-preserving data analyses can be established. It further proposes a concept for a future, self-organising and trusted Web and discusses recommended legal regulations as well as the infrastructure and institutions needed.

1 Ethical and policy issues related with socio-economic data mining

Large-scale data mining is opening up previously unimaginable, new perspectives for science and, of course, even more for business. At the same time, it affects fundamental rights of individuals in ways, which are hard to fully oversee. Among these, the right of privacy is surely one of the most endangered, but it is not the only one. Such risks result not only from single research or data-mining activities. They arise in particular from the combination of singular observations in larger datasets, which contain more and more information, and are capable eventually to depict accurate personal profiles. With these giant data conglomerates at one's disposal, making sense of unpersonalized and apparently irrelevant information is easier than one could think [1]. However, it is still not clear what the implications of developing such *informational cornucopias* are. In the meantime, the construction, enlargement or acquisition of mega data centers run by private companies and national security agencies spreads more and more [2–6]. Intel, the largest CPU manufacturer in the world, has declared that already by 2012 mega data centers will account for 20 to 25 per cent of its server chip sales [7].

In the following sections, we will discuss ethical aspects of building gigantic supercomputing ICT facilities for large-scale data mining, as the ones mentioned before. Our analysis will be primarily based on and guided by a literature review of ethical research in the social sciences. The approach followed can be characterised essentially (but not exclusively) by a positivist and structuralist standpoint, and our discussion will concentrate mainly on privacy issues. However, in section 1.5 we will consider other ethical concerns inherently related to large-scale data-mining activities. Further ethical issues related to social super-computing are addressed in Ref. [8].

1.1 A source-based taxonomy of available personal information

Given that today, more information is available about us than we are usually aware of, let us start the discussion of ethical issues with a picture of the personal data traces almost everyone leaves most of

[1]This is the second part of the paper: D. Helbing and S. Balietti, From social data mining to forecasting socio-economic crises. *EPJ Special Topics* **195**, 3–68 (2011), see http://link.springer.com/content/pdf/10.1140%2Fepjst%2Fe2011-01401-8.pdf.

the time. The following paragraphs provide a non-exhaustive taxonomy of available data organized by data-sources.

Data in public registries

Data belonging this category is generally already public, or available after the payment of small fees to public institutions.

- phone books,
- land registries,
- car plate registries,
- health data,
- salary registries (available primarily for the public sector)
- tax data (public in the US),
- religious confession,
- social security and passport numbers.

Data generated by electronic services

Today, the correct and efficient functioning of our everyday lives is more or less dependent on a few essential services, which are increasingly supported by ICT and electronic infrastructures. This means that, by using such services, a lot of data are automatically generated as by-product. Data in this category are usually available only to certain public institutions and/or some private companies providing these services.

- phone call logs,
- flight passenger information (such as e-mail addresses, credit cards, etc., particularly for flights to the USA),
- bank account data,
- credit card numbers,
- money transactions (e.g. Swift system),
- consumer data ("people who bought X have also bought Y"),
- behavioural analyses.

Data generated by Internet activities

"Look but do not touch" was considered a wise advice to follow when entering unknown environments. However, in the Internet, this is no longer sufficient. The sheer surfing activity, without any content and without accessing any service requiring authentication, e.g. reading a certain news, is enough to generate a wide range of differentiated digital traces. These traces are stored on private remote servers as well as on the local drives. This includes

- Internet service provider logs (e.g. IP and MAC addresses),

- logs of remote access to phones and computers,
- browser history,
- browser cache,
- cookies,
- search queries, and
- click streams.

Data from portable devices

In many social strata, the everyday usage of portable devices is becoming a wide-spread habit. The current integration trend makes portable devices more and more interconnected with each other through wireless communication networks. This facilitates the spatial tracking of persons via location data, which are exchanged by their devices. Such data include

- GSM, UMTS, and GPS location data,
- WLAN/WiFi open hot-spots,
- bluetooth devices,
- RFID data,
- car transponders for automated highway toll payment systems,
- electronic badges (e.g. for conferences [54], hotel rooms, etc.)

Moreover, the large availability of peer-to-peer connections and Internet access points increases the risk of security breaches and data leaks, especially when these devices are used by people unaware of their vulnerabilities.

Finally, the portability of such media introduces the risk of loss of the device itself and consequently of all data stored in them. Given the ongoing miniaturization process and the steady improvements in capacity, the privacy concerns arising from the lack of encryption or other data-protection techniques for such devices are real and concrete. This concerns, in particular,

- video and photo-cameras,
- mobile phones,
- electronic agendas and smart phones,
- laptop,
- flash memory cards and external hard drives, and
- smart multimedia players.

Unauthorized content captured from diverse multimedia devices

Individual actions that reveal the way of living of people may be recorded in both public and private venues and made public at any time and without any previous warning. This concern is increasingly more concrete due to the integration of multimedia contents into global projects such as Street View, and the success of photo and video on-line repositories. This concerns

- uploaded content on social Web sites (e.g. embarrassing party snapshots or videos),
- Google Street View photographs,
- public webcams.

User-generated contents

Many users "voluntarily" share personal opinions or even detailed personal information on their on-line profiles. Whether they are aware of all the risks of this practise it is not entirely clear, but the material is sufficient to identify political, religious and/or sexual preferences of many Internet users. This concerns

- blogosphere data (forums, blogs, chats, etc.),
- the archive of mailing lists or discussion groups,
- keyword scans of free mail accounts,
- social network data.

Security data

Under the flag of security, people were willing or forced to reduce the range of their personal freedoms, with consequences often also for personal privacy. This can happen through an explicit disclosure of personal data, e.g. filling in a security form to enter a foreign country or through accessing a given service, or tacitly, e.g. through public surveillance cameras.

- video surveillance (CCTV),
- face recognition data,
- biometric data,
- audio recordings, directional microphone observations,
- phone call surveillance,
- speed radar photographs,
- scanned items and body scans at airports,
- security forms that must be filled in.

Intercepted data

From very basic to very sophisticated techniques, despite this may be for illicit purposes, electronic communications can be intercepted. Examples include

- network eavesdropping (emails traffic, phone calls, etc.),
- identity theft,
- hardware trojans,
- software trojans,
- the physical analysis of variations in electromagnetic fields of wireless devices (keybord and mouse) and of computer screens,
- the monitoring of fluctuations in the electricity consumption of electronic devices.

While the above lists are probably not complete, it is obvious that the combination of only some of the above data can eliminate privacy to a large extent. Modern information services give a striking picture of this (see e.g. [69]). On the one hand, they show how much information can easily be gained about a single person (contact data, pictures, videos, news, etc.). On the other hand, they illustrate how easily wrong information not related to the person searched for is mixed between correctly retrieved information. Therefore, we will discuss below whether privacy is just an outdated concept, or whether it is crucial for the functioning of democratic societies.

1.2 Why would the honest be interested to hide?

When it comes to private data, some people suggest that privacy is mainly in the interest of dishonest, criminal, or pervert people. In the following, we will argue that this is a dangerous misconception. Privacy has been granted not as a concession of the state to the individual, but because a modern society needs it in order to florish.

Although different in several respects, commercial confidentiality may serve as useful illustration to explain why privacy is needed by individuals. For example, if confidentiality would be dropped, there would be no incentive for companies to invest into expensive long-term innovations, which pay off only through a winning margin. It would be so much cheaper to copy inventions of others as soon as they occur. (There would not be such a fierce discussion about copyright protection/patent enforcement, if this would not be the case.) Secrecy and confidentiality are needed to gain a competitive advantage (in particular in time) that makes innovation commercially profitable. There are two other interesting point about innovation:

- Innovation usually starts off in a minority position [17]. In the beginning, there are a few supporters and customers only. In other words, there is little innovation without the existence of minorities.
- As is known from evolutionary theory, innovation thrives best when there is a large diversity of variants [18]. In other words, diversity or "pluralism" is the motor driving innovation. Would we just orient ourselves at the majority or what is "normal" (the average), the innovation rate and, with this, adaptability to changing (environmental) conditions would be poor. This is actually the reason why totalitarian regimes are sooner or later destined to fail.

These principles can also be transferred to individuals. Without privacy, pluralism is in danger, as the following lab experiment shows [19]: Experimental subjects had to guess the correct answer to a factual question such as "How many murderers occurred in the year 2006?" They received a certain amount of

money, whenever their answer was close enough to the correct one. In one setup, they decided several times without any information feedback, in another setting, they were informed about the estimates of the others. In the first round, the variation of answers was high, but the correct answer was always within the range of answers and was usually well approximated by the geometric mean value of all estimates. When information feedback was provided, the answers converged over time, which may be taken as sign that the right answer was identified. Instead, however, it often happened that the relevant spectrum of answers did not contain the correct answer anymore. In other words, social imitation created herding effects, which were often misleading.[2]

The financial crisis is probably an example for such herding effects, which led to extremely expensive mistakes. Herding-related mistakes would become even more likely, when people were put under pressure to conform with frequent opinions or behaviors, and as the above experiment shows, even when they would only be exposed more often to other opinions than they used to be. This applies as well to many current Web services and recommender systems, which reinforce dominating opinions.

Revealing private data would increase this tendency of conformism enormously and would have other unwanted side effects, as the following points indicate:

1. If behaving "normal" would become the standard and individualism would be discouraged, life would become more predictable, but for sure also much more boring.

2. Conformism implies a danger of discrimination (for having a certain religious belief, age, gender, disease, sexual preference, etc.; it is not without reason that Americans apply for jobs without a birth date and without a photograph). It is well-known that minorities need protection. One must be aware that it is usually minorities who create the concepts and life-styles of tomorrow, and that it is hard to say in advance, which ones will it be. The minority behavior that eventually wins a majority largely depends on environmental changes and historical developments. A society, therefore, needs to have a pool of minorities to successfully adapt to the challenges and opportunities of the future. Minorities are an indispensable ingredient in the process for evolutionary innovation [18, 29].

3. The majority behavior of today may be a minority behavior of tomorrow. What is normal today may be perceived as abnormal tomorrow. For example, it is hard to predict how we will think in the future about the appropriateness of certain kinds of food we eat or the environmental and labor conditions under which goods that we buy are produced. Hence, nobody can be that sure his or her current behavior would be considered proper in the future. Social norms are continuously changing [116]. For example, in the 60ies, the values of society were changing dramatically, and the establishment got under enormous pressure. There are many other examples, such as racial segregation, which was considered "normal" by many people in the past, but is seen in a totally different way today.

4. Private data could be misused by companies. For example, insurance companies have an interest to offer cheap contracts to the majority of people and to charge minorities for special risks (e.g. inborn or past health risks, or higher hospital costs of women giving birth to children). This, however, clearly undermines solidarity.

5. Publicly available data could be misused also for criminal purposes. For example, the city of Oakland releases information on where and when arrests are made, which is later on displayed on a private Web site [70]. From that Web Site, it was possible for criminals to infer the police tactics, patrolling times and other valuable information [30].

[2] Note that taking the wrong decision occurred here even without social pressure, while it is known since the famous Asch experiment that individuals give predominantly wrong answers (against their own judgement), if the people before them do so [20].

6. Companies start charging money to people who want certain private information to be deleted [56–59, 73]. A recent newspaper article even predicts that privacy in future will be a privilege of the rich [104].

7. Disclosing the wealth of people explicitly or implicitly (e.g. through Street View services) can endanger individuals and increases the chance that they may become victims of crime. Therefore, being rich may become less rewarding, and all the private initiative, innovation and commitment leading to it as well.

8. Generally, people with professions that require them to take unpopular decisions sometimes (such as judges, policemen, or teachers) need a certain degree of protection of their private sphere. Otherwise, they will not be able to exercise their job seriously anymore and end up doing what pleases those they have to judge.

9. People may lose the chance of forgiveness of the mistakes they have made, if information about them remains publicly accessible forever [28]. In the past, after a reasonable punishment, depending on the gravity of the misconduct, the policy of societies was to forget about them. In the worst case, delinquents could still emigrate to other countries, where nobody knew them, paying with the abandoning of their hometown a high price for getting a *second chance*. Now, however, wherever one may go, the digital traces left behind will follow him or her. This is not necessarily bad, but it certainly requires a savvier society that is able to remember and forgive at the same time. As Thomas Szasz said "The stupid neither forgive nor forget; the naive forgive and forget; the wise forgive but do not forget." Without an adequate mix of tolerance and solidarity, the ability of a society to (re-)integrate people could be seriously undermined. Outcasts would only have a chance to find friends among other outcasts. As a consequence, this would fragment society into a variety of subsocieties—a tendency, which can be observed already.

10. Whenever a huge amount of personal information is available, individuals, private businesses or public institutions may try to infer individuals behaviors, preferences and attitudes and to classify them according to certain profiles. This tendency is as strong as dangerous, since there is no such thing as an accurate classification. Moreover, in the presence of information asymmetries, which are extremely common in everyday life (such as market exchange, buyer/seller interactions, insurance contracts, bank operations, job interviews, etc.), an inappropriate or wrong classification may be hard to correct or oppose to. Moreover, it may affect the lives of people in manifold and unexpected ways, given the high degree of interconnectedness of different services. In the worst case, it can even drive people through no fault of one's own into *circuli vitiosi*, from which they cannot escape. For example, missing the repayment of the leasing of the car once could mark somebody as "insolvent" to the system. This label would prevent this person from getting future loans, which he or she would need in case of temporarily financial reverse. However, it could lead to even more paradoxical situations. For example, by skipping one installment, the system would automatically register the fact "interruption of contract", and tag one's profile with a negative label. Ironically, the real motivation behind the fact "interruption of contract" could even be that the whole amount of money due was paid at once, without waiting until the contract expired.

    The above example is real, and wrong classifications like these are already happening. But that is not yet the worst possible scenario imaginable. In fact, we must be aware that any form of classification introduces elements for discrimination, because the "labels" are often not fitting and not mutually agreed on [21]. Classifications (be they justified or not) create peer groups and may seriously undermine the basis of cooperation and shared norms in our society. They may also cause unnecessary conflicts [22].

11. As it becomes possible to learn quickly what kind of people we are interacting with and what they do and think, this will undermine an independent judgement of their qualities (and weaknesses).

Rather than giving everybody a fair chance to find the right kind of friends, people might be put into a certain "box" and socially excluded. It is known that people need to be protected from intolerance, mobbing, blackmailing and bribery. To live in peace, people often choose to segregate themselves from others. Given the availability of a lot of personal information to everyone, however, the Internet does not allow this anymore. In this connection, it is important to note that undermining the mechanism of voluntary segregation can seriously affect the cooperation among people, to the disadvantage of everybody [23, 24].

12. The more the Internet knows about everyone, the closer we get to to a situation where we can effectively read other people's minds. Such a situation, however, would potentially generate a lot more conflicts than we have today.

13. It must also be noted that having more information freely available does not necessarily lead to a more transparent, fairer or better society. In an information-rich environment, people spend only a short time on a certain subject, and it easily happens that people get a wrong impression based on such a *pars pro toto* approach (assuming that the first or a randomly picked piece of information would be representative for the full information). Therefore, large amounts of information can promote misjudgements of somebody's behavior by the press and by the public opinion [31]. Such reputational effects are difficult to correct, particularly as rectifications (e.g. when a suspect in a crime case has been found innocent) are often poorly noticed. This may have a serious impacts on individual lives.

14. When everybody has the same information at the same time (and at negligible costs), this may have negative feedback effects such as herding effects. A typical example is the information about a traffic jam, which is provided to everybody via the public news. One can easily imagine that this leads to over-reactions of drivers to the news and, thereby, to overcrowded alternative roads, while the originally congested route may become underutilized. A possible solution of this problem is to provide user-specific information according to probabilistic rules [105] or to overlay randomness to the information signal [106].

15. Systems where a high degree of transparency has already been implemented for years have shown to become more sensitive to sudden regime shifts. Examples are market hysteria and volatility clustering phenomena, which can cause failure avalanches. In some cases, transparency on the producer side can also facilitate the establishment of tacit collusive practises, as it has been found in on-line markets, auctions, and laboratory experiments [10–12].

16. Decisions to reveal private information may even spread in an "epidemic" way. For example, if someone decides to provide access personal data (such as GPS car tracking data, in order to get a cheaper car insurance contract), this can deteriorate the conditions and potentially narrow down the options for those, who do not want to give up their privacy. In other words, revealing ones own data can have an impact on other people who did not like to do so, but who are eventually forced to provide private information in order to maintain the same contract conditions and the same price they had to pay before. This also applies to private health insurances, for example.

17. The data on the servers of certain Internet companies probably know more about us than our friends and partner, and maybe even better than ourselves. However, when knowing the preferences of customers, companies may try to manipulate their choices, and possibilities to do so may increase with personalized recommendations (special offers may even have addictive effects). As it becomes possible to shape the customers expectations, this is likely to decrease the willingness of producers to tailor products and services to the needs that customers really have. In fact, due to the "economies of scale", businesses have a natural interest in providing a number of standard products.

18. Finally, recent scientific studies indicate that pluralism in a society may get lost, as new technologies change the parameters of the opinion formation dynamics [107]. Socio-diversity and its benefits (as outlined above), may easily get lost in favor of conformism and monoculture. It requires the mechanism of individualization, i.e. the desire to be different from others. Therefore, technologies or circumstances promoting conformism may seriously endanger the basis of democracies. In fact, the danger to suppress minority opinions and preferences increases as large datasets containing private information are centrally stored, and as it becomes possible connect different kinds of datasets. It is clear that knowledge implies power, and it would be naïve to think that people would not use it. In fact, there are many examples of misuse of private data (see the section on cyber risks below). It would be surprising, if organized crime would not try to get access to Google's data. One of the few laws of social systems, which have been confirmed again and again is: "Anything that can go wrong ... generally does go wrong sooner or later," This is concerning, as today's information systems probably *would* give someone the power to damage today's pluralistic societies, if he or she really wanted. After all, the Internet contains more sensitive information and about a larger number people than secret services of totalitarian states ever had. In addition, experience tells us that no database is absolutely safe. In 2009, for example, several large sensitive datasets have been stolen from public institutions in Great Britain, where they should have been well protected [25].

Therefore, the storage and processing of large datasets of socio-economic activities is a very sensitive issue. They certainly have the potential to harm pluralistic societies. The interests of individuals (such as privacy) and companies (such as details of their business) *must* be protected. Therefore, it is necessary to address cyber risks and ethical issues by scientific, legal and technological means. The following sections provide guidelines, how this could be done.

1.3 Cyber risks and trust

Big data aggregates represent much sought-after targets for cyber criminals and big challenges for security experts. The Symantec Internet Security Threat Report XV [32] mentions a 100% increase in the number of new malicious programs identified (more than 240 million in 2009) and estimates the number of Internet users (companies and individuals), who have been victims of cyber-attacks trying to steal money or confidential information, to be of the order of 360 millions. More and more attacks are aiming at *identity theft*. Sixty percent of all data breaches that revealed identities were in fact the result of hacking.

An incomplete list of the risks one must be aware of when using the Internet today is given below:

- data theft,
- theft of pin codes and passwords,
- identity theft,
- viruses, worms, and trojan horses (damaging software, steeling passwords, etc.),
- data manipulation,
- wrong evidence (wrong accusations),
- damaging rumors [62],
- information pollution,
- spam and unwanted advertisements.

These risks may seriously undermine the trust of people in the Internet and services based on it. For example, the theft of access data for electronic banking through phishing attacks has recently become a wide-spread problem. However, trust is essential for economic exchange. Systems which would effectively not work without a certain level of trust include:

- electronic banking,
- e-mail,
- eBusiness,
- eGovernance, and
- social networking.

To solve the above problems, the right mixture between legal regulations and technical innovations is needed.

1.4 Current and future threats to privacy

Whether personal data disclosure in the Internet is the result of a truly *voluntarily and deliberate* choice is rather questionable. In social research, voluntarily participation is considered a basic human right, which overlaps considerably with the principle of informed consent [40]. Moreover, European law, for example, gives individuals an individual right of control over personal information.

There is no unanimous definition for informed consent, but according to Diener and Crandall [26] it is "the procedure in which individuals choose whether to participate in an investigation after being informed of the facts that would be likely to influence their decision". In principle, any decision can be considered as informed consent if it has been taken after being provided with the amount of information that *a reasonable and prudent person would want to know* [108]. In the Internet this is seldom the case. In fact, it is both possible and relatively common for individuals to access Web sites without reading the terms and conditions (which may be several dozen pages long). It it also unlikely that most people would understand the full contract, while they actually have to approve this. Moreover, they are usually not given any options rather than accepting the conditions in order to get the requested service or rejecting them at the cost of no service, which does not give users a reasonable choice. Under these circumstances, people may nominally give consent, but without being fully aware of or agreeing with the terms and conditions. Such a situation would not be considered as informed consent [41]. This contravenes a widely accepted principle in Social Science Ethics that states that "*as far as possible, participation in sociological research should be based on the freely given informed consent of those studied*" [39, 109]. Moreover, fully informing the respondents it is not yet enough, since researchers should endeavour to make sure that the participants of an experiment have fully understood risks and consequences [108]. This applies in particular for physically or mentally challenged individuals [109, 112], but cannot be ensured in the Internet [41].

Whether large data-mining companies are aware of the above mentioned ethical issues is questionable, especially when CEO's of big data mining companies make statements about privacy such as the following one: "*If you have something that you don't want anyone to know, maybe you shouldn't be doing it in the first place*" [110]. This is worrying, because if ethical standards turned out to be insufficient at some of the fundamental places of command of the biggest data-mining companies, or if market competition would push them to pursue only the logic of profit, what would refrain them from collecting and using people's data even in illicit ways? Data mining techniques improve every day, while regulations and control over the gathered data are lacking far behind. For example, tracking the source of collected information—once it is stored in secured and not publicly accessible databases—is virtually impossible; knowing who has access to which kind of personal data is also not possible today. Relevant to this discussion and particularly

controversial is the latest case of Street View cars. For several months, these cars have been storing personal data, including passwords, credit card information and accessed email contents, which were intercepted from private WiFi networks. The incident was reported as result of a programming error, but others have suggested that this was rather a case of WiFi sniffing, as there exists a software patent which involves intercepting data and analyzing the timing of transmission as part of the method for pinpointing user locations. At the time of writing this White Paper, the actual situation is still unclear [35–38].

Also when not possessing a sophisticated and expensive data mining system, criminals can collect illicit data easily through Web browsers, as these are daily affected by new malicious exploits (see [86]). The most common attacks are now based on a technique called "history stealing". Some Web sites even show this security issue to visitors [75, 76], thereby demonstrating how easy it is to extract personal surfing habits of Internet users. Scientific literature on the topic is vast, and latest studies conducted on 243,068 users found that 76% of them were vulnerable to history detection by malicious Web Sites. Newer browsers such as Safari and Chrome were even more affected, with 82% and 94% of vulnerable users [33]. Unfortunately, there is yet another privacy issue related to recent generations Web browsers: their inherently high customization capabilities have made them unique, and therefore trackable. In fact, even disabling cookies, and blocking history stealing-like exploits, individual Web surfing can still be reconstructed by simply following the customized "fingerprint", which the browser is carrying around from site to site. This fingerprint is actually made up by all the configuration information that the browser is exposing to remote Web sites. According to the Electronic Frontier Foundation [77], information such as which plugins are installed, which fonts are available and which operating system the browser is running on, can create a unique portrait of 94% of the visitors (for a self-demo see Ref. [78]).

Unethical or dishonest intents are not the only pitfalls glooming over on-line data sharing. Even in a scenario, in which one has consciously provided his or her own personal data to a company, which is using them lawfully, unforeseen issues can suddenly arise. For example, such a company could be sold or merged with another one, or simply, the data could be sold, based on a change in the data handling policies. Users are typically not notified of such changes, and they usually have no effective possibility to draw back their data and their consent to use them. Some social network Web pages are examples for this. In fact, because of the continuous updating and modification of the terms of use [43], the Electronic Privacy Information Center (EPIC) has filed a formal complaint at the US Federal Trade Commission [44], and more lately US senator Charles Schumer (D-N.Y.) has petitioned the Federal Trade Commission to request that the agency addresses the issue of social network privacy policies [42]. Moreover, some national data protection commissioners have publicly warned of using certain social network sites [52, 53]. Just recently, the vulnerability of these services has been demonstrated by someone, who downloaded 100 million user profiles and made them publicly available for download [111].

Joining groups within social networks can offer another exploit for potentially malicious de-anonymization attacks. A recent paper [34] proved that 42% of users that use groups can be uniquely identified. This results are noteworthy, because traditional privacy attacks were based on aggregating information from multiple datasets. Such methods were based on collaborative filtering [46] and enabled an efficient and highly reliable characterization of a person from a few data. The underlying technology is quickly advancing [47], and it may give service providers, such as mobile phone, Internet television, or social gaming centers an unprecedented amount of personal information. Research on related privacy issues and their potential explicit or implicit consequences is still in its infancy [48]. Moreover, an efficient legal protection seems to be missing, while a simple-to-establish solution to some of the above problems would be accountable pseudonyms [49].

Additional risks for the privacy of users emerge, when companies are forced to reveal private data to governments or legal institutions. Google offers a picture of the quantity of data which is handed out to governments [71]. There are also joints startup companies with the CIA [45]. Finally, when data are not subpoenaed or stolen from cyber-criminals outside of the company, they can be leaked in the most fanciful ways, which go from mislaying a physical device containing sensitive information, to the dishonest

action of a single employee from inside the company [50, 51].

1.5 Additional ethical concerns

Ethical problems are intrinsically "ambiguous, uncertain and prone to inevitable disagreement" [112], i.e the correct answer cannot be deduced algorithmically from general rules to particular claims. They are related to cultural values and social norms. In the following, we raise a number of open ethical questions connected with large-scale data-mining activities, to which, of course, we cannot provide definite answers here. Related research programs seem therefore in place. For the time being, governments and companies engaged in large-scale data-mining are advised to follow the procedural ethics approach presented in section 1.6.

- As large-scale data-mining activities are increasingly successful in predicting (aspects of) individual behavior [89], they will constitute an extremely powerful tool. This raises issues of the possibility of misusing it. More importantly, it raises the question of who gets to use these tools on what grounds. Will it be national governments and international corporations? Would there be a moral imperative to make the systems available to developing countries, NGOs etc.?
- What about competing claims of systems? If an early warning system recommends certain activities, how should societies respond to such recommendations? For example, how to handle situations, in which a scarcity of resources occurs?
- Who will own the algorithms and the outcomes of the data mining activities? Intellectual property is often discussed in terms of ownership of data used for input, but the more interesting question would seem to be: who owns the predictions? As they could potentially be subject to patent protection for computer programs and business methods, a rigorous analysis of the implications of intellectual property protection for data mining activities is needed.
- If policy is based on predictions, how open is the system to critical review? Who will know and understand the algorithms? How can mistakes in algorithms be identified and rectified?

1.6 How to address ethical issues in large scale social data mining

Large-scale data-mining raises both procedural and substantive ethical issues. Some of the latter are predictable and solvable by implementing legislative and technical solutions. In the case of privacy, for example, this would include

- the use of scanners for viruses, trojan horses, etc.,
- encryption,
- fragmentation of data [80],
- restriction of access/read/write/execution rights (depending on the type of data and purpose),
- selecting higher security standards in the browser (for example, turning off cookies or deleting the browser history),
- anonymous surfing [81, 82],
- use of pseudonyms.

Nonetheless, one needs to underline that a full understanding of substantive ethical issues would require a full knowledge of uses and applications of the system, which is impossible to acquire a priori. In order to ensure a future-oriented approach to ethics, every project performing large scale data mining should therefore incorporate procedures that will allow the identification of substantive ethics as well as ways of addressing them. Such procedures should include the governance of the project from inception to delivery and cover governance recommendations for the individual components (early warning systems). It should incorporate reflexivity in the project team, continuously discussing the following questions (and regularly seeking independent feedback from outside):

- What are the substantive ethical issues that can be foreseen at any given point in time?
- What are the assumptions underlying the project itself as well as those underlying the ethical analysis (what is perceived to be an ethical issue, and why?)
- How can appropriate processes be established to address known ethical problems (e.g. informed consent procedures)?
- How can factual knowledge about the product and its likely consequences be gained?
- Who are the stakeholders affected by the system and how can their local knowledge be fed into the reflective process?

2 Towards privacy-preserving data analyses

Privacy concerns, although often justified, can cause serious obstacles to socio-economic data mining, while in many cases such data-mining would be in the public interest, when done in a privacy-respecting way. For example, socio-economic data mining would be needed to gain a better understanding of socio-economic problems, how they arise and how they can be addressed. Therefore, the following sections elaborate concepts, how data mining could be done in a privacy-respecting way.

2.1 Deliberate participation

The simplest possibility to do social data mining is to do it with data that individuals share deliberately. For example, some Web sites, such as Blippy.com, Skimble.com or Swipely.com, collect everything from consumer data over the last movie you have seen up how many push-ups you have done in your last training session. Participants of these Web services intentionally make their data available to everybody, and they can be analyzed in any possible way. The only concern from a statistical point of view is that the set of people participating in these Web2.0 activities is not representative for the whole population, i.e. one would need to make complementary analyses in order to learn, how it is possible to correct for biases in these data. Typically, participants are younger than average and are not concerned to share their data because they lead pretty much "average lives".

Further data can, in principle, be analyzed by crawling the Web. The data out there are usually traces of, for example, shopping activities at eBusiness platforms or social networking activities. They are accessible to everybody in small numbers, and it is not clear whether and how much people would care about a company or person analyzing these data in large amounts, as they can be gathered by automated programs such as "spiders". There are certainly problematic applications of this kind, in particular when the resulting datasets are used to do business, although the data were not intended for this, or if they are sold to third parties with unknown intentions.

As the recent discussions about the activities of large data mining companies shows, legal regulations against unauthorised processing of individual data are urgently required. Scientific analyses, which lead to discoveries of public interest, may have a better justification, but it must nevertheless be decided in each

single case, whether individual rights are touched and what is the public benefit of such analyses. Shear curiosity and the publication of a scientific paper may not be a sufficient justification, and therefore, the consultation of an ethical committee seems appropriate.

As a consequence, it would be much better to work with data that people provide intentionally for a given purpose. Statistical samples can already be quite useful. Special "on-demand-data-gathering" tools could allow people to easily opt-in and opt-out of data-collection programs in a situation-specific way. For example, while people may usually object to provide their data, it is likely that the participation rate increases in special situations such as crises, where people tend to change their priorities and make a contribution. However, it is fundamental that the gathered data will be used only for the purpose people have explicitly given consent to. With the project "Gaydar" [103], the MIT demonstrated how easy it is to filter out from publicly available data sensitive personal information, which may be misused. This study predicted the sexual orientation of Facebook users by analyzing the publicly accessible pictures of their friends. Such studies suggest that the processing of data should be allowed only for a certain time period and for the purpose they have been provided, requiring that users have adhered to an explicit, fair, and informed opt-in procedure. For sensitive data-mining activities it would be appropriate to apply the standards followed in clinical studies today. In order to support on-demand participation, particular trust-worthy Internet platforms should enable the case-wise sharing of personal data according to the specified purposes. This could be a special function of future eGovernance platforms.

2.2 Anonymization and randomization

To satisfy the data protection directive 95/46/EC, any data containing personal information needs to be anonymized before it is evaluated. While this may be sufficient for many simple analyses, it may not guarantee that the identity of individuals cannot be revealed from anonymized datasets. Substantial research has been and is currently being performed in the database community on privacy preserving data mining, reflecting the importance of this subject [93–97] (for a comprehensive state-of-the-art summary see the "Privacy-Preserving Data Publishing" survey [98]). Nevertheless, there are still a number of open problems, and many approaches are standing next to each other, lacking user-friendliness, integration, and a consequent systemic approach. Problems occur in particular when datasets contain a list of many different features, and some combinations of features are rare. As a consequence, such data must be sufficiently coarse-grained and/or randomized to make sure that combinations of features occur in sufficiently large numbers and cannot be individually resolved. Furthermore, it must be avoided to save lists with many features in one single dataset. It is safer to store them separately on different computers and to access the separate datasets only with programs, which are guaranteed to determine coarse-grained properties only such as (sufficiently rough) statistical distributions. The resulting derivative datasets should be comparatively small and unspecific, or they should be surrogate datasets, in which the relevant *statistical* properties are the same, but the underlying individuals (persons, companies, etc.) are randomly reshuffled and not identifiable anymore.

The generation of the anonymized, derivative, and surrogate datasets for the original data should be done by particularly qualified and trustable institutions, while a larger number of people can work with the resulting, less critical datasets. In the last decade, research in privacy-preserving data analyses has produced methods and tools aimed at publishing data under a privacy-preserving shield. For example, data are made anonymous with respect to a certified trustable anonymity notion, which essentially guarantees that the probability of tracing back any data to the identity of the person to whom the data originally belongs is so low that it can be considered null in practice. Another active research line concerns the privacy issues in case of mobility data such as those produced by location aware devices [101, 102].

To protect the original datasets from theft and unauthorized access, the specially secured and authorized data centers should store them in an encrypted way, and decryption should be done only piecewise and for the miliseconds, when the derivative data are generated. All commands and source codes of computer programs involved in sensitive operations should be automatically protocoled on a separate server,

which is unaccessible to persons who are authorized to deal with original datasets.

2.3 Coarse-graining, hierarchical sampling, and recommender systems

As indicated before, in case of sensitive data (such as pregnancy, religious confession, diseases or the sexual orientation), it must be ensured that individuals and group memberships cannot be identified from socio-economic datasets. For this reason, datasets for statistical analyses must be coarse-grained in a suitable way. This may also be done by real-time data-mining ("reality mining") approaches, if they are suitably designed. For example, to determine congestion on a freeway, it is possible to analyze mobile phone usage data, but it is not at all necessary to know who is calling whom and what is the content. The same applies to GPS localization information of mobile phones, if the distribution of people is determined for the sake of an efficient evacuation. It is just necessary to make sure that any potentially sensitive data (such as the underlying phone number) is thrown away before the statistical evaluation is performed. However, as the recent case of WiFi recordings by Street View cars has shown, transparency is needed for such applications, as one needs to make sure that really no sensitive data are stored. In principle, it could be legally required that the underlying algorithms are published, and no algorithms may be used which are not open source.

One particular approach in reality-mining could be a hierarchical sampling via ad-hoc networks of, for example, sensors or mobile phones, where detailed information is only processed locally, and any transmitted information undergoes a certain level of aggregation. That is, as data are distributed over larger distances, they undergo several aggregation steps, which may be imagined like a hierarchical sampling method. Whoever wants to process a large dataset, would only get a coarse-grained view of the data, since they would be accessible only via a high level in the data-processing hierarchy. Whoever managed to see data on a lower and, therefore, detailed level, would only have a very short-sighted and limited view, i.e. see very little. It appears, however, that the technical details of such systems matter in order to be sufficiently privacy-protecting and acceptable to users of the resulting services (e.g. location-based ones). A transparency of the data processing algorithms and related legal regulations appear to be needed. It should be explicitly forbidden and prevented to collect and store low-aggregation-level data. It must be ensured that they are deleted directly after they have been processed and before they are transmitted. To be uncritical and widely acceptable, the processing should happen in the technical devices used by the individuals and not on company-owned infrastructures (as it is common today).

A possibility to make low-level data robust to interception would look as follows: Given that the data of interest can be represented as points in a (quasi-)continuous space, one could add random numbers according to a certain statistical distribution. Rather than transmitting the correct value (such as the exact location of the individual), a random number ("noise") would be added, before the value is transmitted to the ad-hoc network performing the reality mining. Such random falsification would make low-level-aggregated data useless and create a "foggy" situation that protects the individual from being revealed [55].

However, if done in a suitable way, the aggregation of the individual data could still lead to reasonably accurate results due to the law of large numbers, according to which errors average out in a statistical sense.

Services of *recommender systems,* of course, need to target an individual specifically, which seems incompatible with overlaying noise. However, recommender systems could still be realized by applying a two-component strategy: The first component would be a rough search, which does not consider individual information or preferences (or only, when sufficient noise is overlayed). Among the search results, the personal computer or smart-phone of the user would then select the individually fitting search hits, products, or advertisements, based on personal information and preferences that are exclusively stored on the individual computer rather than on a system of servers. Putting it differently, recommender systems should be changed from an approach, where individually customized recommendations are pushed to the user, to a pull approach, where the user selects in confidence one option out of a larger spectrum

of downloaded recommendations in a way that does not reveal his or her preferences. Individuals who are even concerned about storing personal information and preference data on their own computational device should have the possibility to turn off the second component, which would then result in untargeted research results and in recommendations, which would not be individually customized.

The same approach can be used in connection with location-based services, the great comfort of which many people do not want to miss anymore. Let us assume somebody wants to be guided to a erotic shop, but does not want the guiding company to know this. The person would go to the center of town, and based on his or her falsified, approximate position, the GPS location service would forward to the mobile phone information about shops in the area. The user could then select among these according to categories, but the selection would only be known to his or her phone. It would not be forwarded to the content provider, nor would the exact location of the user be known to the provider.

2.4 Multi-player on-line-games, pseudonyms, and virtual identities

Another possibility to study social interactions are offered by multi-player on-line games such as Second Life. The advantage of these games is that players can participate under pseudonyms, without revealing their real identity. From an experimental point of view, this has some side effects, as people may behave differently under anonymous conditions as compared to conditions with face-to-face interactions. Still, these effects may be compensated for, and there are a number of behaviors, which come out quite realistically. For such reasons, studying interactions in multi-player on-line games is becoming a research technique, which is used complementary to lab and Web experiments [113, 114].

Some of the artifacts of studying multi-player on-line games result from the following facts (here we assume that the system would not allow the registration of several *identical* pseudonyms):

1. People may change identities, i.e. register as a new user if their previous behavior is sanctioned by other players or by the system ("whitewashing").

2. People may use multiple identities, potentially also in parallel.

To overcome these problems, the following measures can be taken:

- Everybody could get a unique virtual identity, which would be needed to create unique pseudonyms.

- Registering a new identity could be made very time-consuming or costly.

- People may be allowed to join a multi-player on-line community by invitation only (and there would be separate lists of members and pseudonyms, which would be secret and encrypted).

An additional problem is that people may buy an identity (pseudonym) with high reputation or scores from somebody else. This problem may be addressed by performing behavioral consistency checks to reveal the use of the same identity by different people. Alternatively or complementary, the matching of pseudonyms with the unique virtual identity could be sporadically checked (by requiring to enter it).

A unique virtual identity can be generated by a trustable public institution such as the registration office. It is practically an electronic signature that can be used to submit documents such as tax declarations or payments. Note that there are already private companies offering trusted virtual identities/electronic signatures, among which Verisign, GeoTrust and Thawte.

The unique virtual identity would have a finite validity (i.e. it would have to be regularly renewed), and plausibility checks for identity thefts would be made, to invalidate stolen identities (such as for credit or debit cards). The identities could, for example, be generated as follows (where the system would log which administrative person handed out what card): When asking for the virtual identity, a box would be ticked in the files of the respective person, indicating that a unique identity card has been handed out. The identification number of this card would be randomly generated, and the receipt of this card

would be confirmed with a signature, showing a valid photo document (ID or passport). The identification number, however, would *not* be known to the office handing out the card.

To reveal the real identity behind a virtual identity in case of a severe crime, this should require the simultaneous agreement of several independent authorities (e.g. judges [who could also be from trusted non-profit organizations]). Only by combining the keys of two thirds of the respective responsible authorities, it would be possible to reveal the real identity. This can be ensured by saving bits of the identity code in different databases, all in an encrypted way. The access keys and responsible persons controlling these keys would be regularly replaced by new ones in order to avoid corruption. It would also be good to let computers randomly decide, which ones of a number of authorized persons would have to decide whether to reveal the identity or not. This would minimize external influence on the decisions of the respective responsible authorities.

2.5 Anonymous lab experiments

Social behavior can also be studied in lab experiments [105]. In these experiments, it may be desired to ensure anonymity of the participants, as they may otherwise not reveal their true opinions or their normal behaviors. In such experiments, it may be needed that the experimental subjects do not meet the experimenter, and maybe not even meet other experimental subjects.

There are different ways of implementing such a design. For example, individuals randomly passing by an information stand could be invited to participate in the experiment. If they were willing to participate, they would draw a lot with a unique number, and they would enter the number of the lot into a time table, which is hidden from the experimenter. At the time of the experiment, the experimental subjects would show up in separate rooms, where they take their decisions in isolation, based only on information coming from a computer screen. Their decisions would then be transferred via Internet to the other experimental subjects. After the experiment, the experimental subjects would get an envelope with their compensation, which would be pushed into their rooms through small slits under their locked room doors. The subjects could leave their rooms two minutes later. The experimental setup would ensure that nobody would know, who participated in the experiments, and it would be unlikely that the same person would participate several times. Nevertheless, participants may be suspicious whether this setup will really be executed in an anonymous way or whether there is a chance of hidden observation, and this may still affect their behavior.

A similar and even more privacy-protecting setup can be realized via a Web experiment. A large number of people would be informed that the experiment takes place at a certain time and could log on with pseudonyms. Among the people who have logged on the experimental Web page, the computer would randomly match individuals to form experimental groups. At the end of the experiment, each individual would get a voucher with a unique code, which can be exchanged for the compensation for participating in the experiment. One of the following ways of payment may be chosen (listed in increasing order of anonymity):

1. The experimental subject gets the compensation from an independent cashier (e.g. the university cashier) against showing the voucher, without the need to sign a receipt.

2. The person gets the money from an independent, i.e. trusted third-party payment service (e.g. bank or post), when presenting the voucher (i.e. the voucher would basically be a cheque). For example, it would be possible to use the mechanical turk [72] for a third-party payment.

3. The experimental subject gets the compensation by entering the unique code of the voucher into a special cash machine.

Experimental subjects could be recruited in different ways: The simplest would be to display posters in public areas, calling for participation at a specified time via a certain Web page (and people could actually

participate from a computer in a computer pool or Internet cafe, if this gives them a better feeling of anonymity). Similarly, the announcement could be made via an advertisement on a heavily frequented Web portal. At the specified time, an algorithm would match visitors of the Web Site in a random way and try to make sure that groups of friends could not play with each other.

Avoiding that certain subjects participate in the same experiment multiple times is more difficult (at least as long as only a few people have unique virtual identities). One possibility would be to send out invitations to a large number of e-mails, making sure that there is only one e-mail address per person. People willing to participate would enter into a Web page their e-mail address or a unique code sent with the invitation e-mail. This is required for authorization, to prevent multiple access. After this, they would be redirected to a Web page, which shows a large list of unique codes, one of which can be randomly chosen by clicking on it. This will cancel it from the list and inform a Web service hosted by an independent, third party (e.g. a computer center) that this code has been authorized. When the participant enters the code into the Web page of the independent Web service, another code is returned, which is randomly selected from a long list of unique codes. That code will be needed to get access to the experimental platform at a later point.

The above procedure makes sure that the first step prevents multiple access. Afterwards, the selection of an individual code makes sure that the third party cannot have any clue of the relation between this chosen code and the e-mail address of the experimental subject. While it knows the list of acceptable codes, it does not know the identity of the person, just the fact that it is authorized to get a randomly chosen code from a list of unique codes, which are accepted by the experimenter. However, which code is randomly selected by the computer of the third party cannot be known by the experimenter. Finally, any temporal correlation among individual registrations is lost by implementing a sufficient time delay, after which the actual Web experiment takes place.

3 Concept of a future, self-organizing and trusted Web

In the following, we will describe technologies, which give people back control over the data available about them. Some of the following runs under the label of privacy enhancing technologies (PET). For example, most Web browsers today allow one to turn off cookies (which, however, makes certain Web services disfunctional). Furthermore, there are tools such as Tor [81] and Freenet [82], which support anonymous Web browsing and anonymized content sharing by obfuscating the IP address of a computer. However, this is still not sufficient to guarantee anonymous web browsing [78,79]. Furthermore, one serious problem of today's Internet still is the fact that it does not forget and that it does not provide control over copies of data, which somebody has uploaded in the past (e.g. party photos). First solutions for data with finite lifetimes have become available only very recently [27].

3.1 Data format

The following concept of a future, self-organizing and trusted Web is aimed at overcoming the above mentioned and other problems. The basic feature of the concept is a new "Helbietti" file format, which electronically signs and encrypts contents, but has a number of unencrypted specifiers such as

1. a unique file identifier (which is different for copies),
2. the kind of content (factual information, advertisement, opinion, unspecified),
3. the lifetime (from ... to ...),
4. a public annotation field allowing to tag the file and link it to others or to link it to comments or ratings, and

5. information regarding the price of producing and receiving a copy of the file.

There would also be encrypted specifiers (readable only to authorized users), such as

- the originator of the data (anonymous, pseudonym, real name, or company name),
- the owner (anonymous, pseudonym, real name, or company name),
- the date and time of generation,
- a unique content identifier (e.g. check sum),
- locations of authorized copies, and
- the persons or groups authorized to read, modify, or execute the file (which would again be based on real names or pseudonyms etc., but one possible specification would be "everybody").
- Annotations, which could be read only by the authorized persons or groups.

The following data would be double encrypted and accessible only to the owner of the file (and jointly to a specially authorized group of inspectors, see below):

- the file identifyer(s) of the file(s) it has been derived from (i.e. the previous version(s), if one existed, otherwise null) and the files that have been derived from it (e.g. any identical or modified copies),
- all information regarding money transfers between customers or users of a file and the owner of its content as well as the respective tax authority, and
- the digital rights management settings (e.g. maximum number of copies that can be made from the original file).

To ensure that privacy and intellectual property rights are not undermined, checksum error-detection techniques would immediately reveal unauthorized manipulations to the original copy. Semi-automatic filtering measures could be implemented on servers which would refuse storage and forwarding of tampered copies. This kind of filtering may be compared to the immune defence system of the body against harmful viruses etc. For issues of copyright protection see Secs. 3.2 and 3. Moreover, depending on the sensitivity of the data (public, restricted, confidential, secret, etc.), they would be fragmented and distributed over several files stored in different locations [80]) and additionally password-protected, potentially requiring several passwords from independent authorized persons to access them.

3.1.1 Finite life-time data that can be controlled

This concept immediately allows one to limit the life time of data, as they could only be decrypted within the specified time period. (Although there are first software solutions in this direction [27, 74], they seem to require further enhancements.) In order to avoid tricking the file by modifying the time on a particular computer, the file would automatically have to verify the time with one or several randomly chosen, trusted servers (depending on the level of confidentiality; of course, there would be a long list of such servers). Additionally, the file could be opened in this time window only by individuals or groups that are listed as authorized.

Besides, one could foresee a further restriction to the access of a file by requiring that either the original file or one of the authorized copies are still accessible somewhere in the Internet. That is, if the owner of the file would delete the original file and the authorized copies he or she may have created as backups, no copies of the file may be opened any longer. This would give the owner of the file perfect control over its distribution—a fact which is also important for copyright protection (see Sec. 3.2).

3.2 Intellectual property rights

The new data format also provides new possibilities to protect copyrights better. As music or video files would be encoded and require a certain password to open it, access to the file could be restricted to a single user or group of users. Moreover, Helbietti-formatted files could be set up in a way that a certain prize is charged (e.g. to a prepaid account) whenever a copy is produced. During the copy process, this amount of money would automatically be transferred to the owner of the intellectual property rights, the intermediate seller (e.g. a shop or the person whose file is copied) and the respective tax authorities. This would facilitate a "viral marketing", where users are distributors, who can earn money by disseminating file contents, while benefiting the owner of the intellectual property rights. This would, of course, not prevent the filming of videos and the illegal distribution of related copies. However, this problem could be minimized by a combination of the following measures:

- using pricing schemes that people consider fair,
- selling copies of different quality at different prices,
- allowing users to download contents with pseudonyms and anonymous payment services (e.g. [73]), such that providers cannot track which contents are bought by what customers.

Massive copyright violations could be reduced by using the labeling, reputation and sanctioning mechanisms described in Sec. 3. Also note that the proposed file format allows one to make all copies unaccessible by deleting the single file that the copies were derived from (see Sec. 3.1.1). Finally, for serious cases of piracy the new file format provides a possibility to track from what file a copy was derived, if decryption has been decided by a number of specially authorized people (see Sec. 3.3).

3.3 Trust management

3.3.1 Rating and reputation

Many public goods such as reliable information systems are very difficult to create and easy to exploit and/or destroy. This creates dangers for the quality of "the commons" (public goods). In absence of clear responsibilities, such as it is often the case in the Internet, large collaborative efforts are not encouraged. In fact, contributors are more difficult to identify and to reward, while vandals and other detractors can easily thrive.

Therefore, the self-control of the Web, based on suitable reputation concepts, would be a desirable feature. In principle, people should be able to rate, tag and comment on any data they have accessed. Also ratings and comments could be rated, which would earn the rater a certain reputation. Ratings would not be one-dimensional, but done on a multi-dimensional scale (which could be customized in a user-specific way). The multi-dimensionality is important to support pluralistic, community-specific views.

Note that details of the design of the rating mechanism are crucial. Manipulations of ratings must be prevented. The rating of the raters can serve this purpose, if well constructed. It determines their weight in the calculation of an aggregate rating. The design should be able to distinguish votes coming from robots and from humans. Furthermore, whitewashing (a new pseudonym) and sybil attacks (the creation of many pseudonyms) should be prevented (see the previous section regarding possible ways to do this). Furthermore, to disclose a manipulation of the own reputation via pseudonyms one is controlling (or a mutual manipulation through a friendship network), consistency checks will be needed. That is, at random times, it will be necessary to compare the reputation values that a pseudonym has from the point of view of several others (randomly chosen interaction partners, also independent outsiders). This comparison of reputation values is something like a gossip strategy. If the values are sufficiently consistent, everything is fine and the reputation seems reliable. Otherwise, there are reasons to be suspicious. In such a case, the pseudonym would be labeled for the purpose of intensified observation in order to reveal the

manipulation. Such a differentiated inspection strategy, which focuses on indviduals with a suspicious reputation (and newcomers), but which otherwise restricts to random inspections, saves computational resources but can keep the level of fraud low.

Furthermore, the contents that users upload in the Internet would be rated by other users who have access to them, earning the provider of the content a certain reputation. This offers a tool to separate high-quality from low-quality contents. In order to avoid opinion dictatorship by the majority and ensure socio-diversity (pluralism), it will be necessary to allow for community-specific and multi-criterial ratings. Communities would either result from social networks, or they could be determined via community detection algorithms, identifying groups of people with similar rating, tagging, and commenting habits, i.e. with similar preferences and tastes (so-called "quality collectives" [83]). It should be remembered here that the identities of the people belonging to a community will usually not be known, but rather be composed of virtual identities, namely when pseudonyms are used.

The community-specific ratings, tagging and comments can serve to create filters for certain contents. Therefore, it is possible to design community-specific recommender systems which prioritize contents fitting a community's or an individual's taste. Similarly, undesired contents can be excluded so that it becomes possible, for example, to protect children from sexually explicit or violent contents. In other words, users could tag illegal or inappropriate contents. In serious cases, this could trigger sanctions (see below) or even legal action. For instance, the access to the file could be restricted (e.g. to people above a certain age), or the decryption could be disabled by a certain code foreseen by the cryptographic algorithm. Also, in case there is evidence that access to certain encrypted contents is in the public interest (e.g. relevant for public security), the encryption method could foresee a decryption. In order to avoid misuse such as censorship or violation of privacy, both, the decision to restrict access and to enforce decryption of a file or list of files, would need a certain number of randomly selected, generally trusted and authorized people to agree on the action that needs to be taken. Consequently, such actions would require the application of several keys at the same time. To avoid unjustified decryption by bribing authorized people, these should be replaced after a certain time period, which means that the keys unlocking a file need to change or be changed over time.

3.3.2 Sanctioning mechanisms

In reality, a reputation is hard to earn, but easy to lose. This suggests that, besides a reputation mechanism, the self-organizing Internet could foresee certain sanctioning mechanisms to facilitation a high level of quality. Sanctions may include everything from low ratings, over certain kinds of tags and critical comments, up to banning specific contents within a certain user community. Particularly destructive behavior may be sanctioned by temporary bandwidth reduction. For instance, manipulating ratings or reputation values by sybil attacks (self-ratings via multiple pseudonyms) should be sanctioned in one way or another. The same applies to wrong declarations (e.g. labeling advertisements or opinions as information). People should have a freedom to express their opinions, but they also need to have a chance to distinguish opinions from facts. Furthermore, spamming the Internet with low-quality contents should be sanctioned. Note, however, that what constitutes low-quality content for one community could constitute high-quality content for another community. That is, the sanctioning would usually be community-specific. Only in exceptional cases would it be generally applied.

3.4 Microcredits and micropayments

It also seems wise to foresee in the future Internet the possibility to collect microcredits for small contributions to the public good "Internet". Such microcredits would allow one to reward people, for example, for contributions to public encyclopedias or also for rating contributions or reviewing (commenting on) them.

The data format of microcredits would, therefore, not only contain a certain value ("number of points"). It would potentially also contain (usually in a sufficiently anonymized or encrypted way) information about who owns it and what it was earned for or paid for. Moreover, it would be a tradeable unit, which could contain pointers to who owned it last and whom it is being paid to (again in an encrypted way). Having both backward and forward pointers supports double book-keeping when needed. In mathematical terms, rather than a being scalar (which implies a number of fundamental problems), a microcredit would be an element of a microcredit network connecting values with pseudonyms and merits or items bought. These elements would have a certain number of weighted links (in-degrees and out-degrees) reflecting cash flows. Therefore, it would be possible, in principle, to distinguish different kinds of currencies for different kinds of contributions, and it would also be feasible to a certain extent to analyze flows of microcredits between pseudonyms over time in a privacy-respecting and confidentiality-protecting way (see the section on reality mining regarding how to do this; note that companies could use different pseudonyms for different organizational units, and that they may change them over time). Such kinds of analyses would be enormously useful to determine instabilities in the microcredit market. It would also be possible to give money a history and, therefore, distinguish "dirty money" (such as "blood diamonds") from ethical investments, as certain customers demand them today.

3.5 Transparent Terms of Service

In order to sign-up to a service in the Internet, one is more and more often placed in front of a long list of obligations and contractual clauses applicable for any sort of special case, which an ordinary user has not the adequate competency nor the necessary time to understand. The result is that they are skipped and blindly accepted. Based on such "acceptance", companies can grant themselves a great freedom of action in handling the personal data of their users. This should not be allowed, and large-scale data mining activities should be protocoled and publicly controlled.

Anybody willing to start collecting data from the Internet, or other private and public nets should first publicly provide a legally binding declaration about what is done with the data and why. In particular, it should contain whom (what companies, institutions, etc.) the data will be shared with, and what is done with them exactly.

This declaration should also comply with an international "*data-collecting protocol*", which needs to be established to set legal and ethical constraints to the action of *data harvesters*. The protocol should define minimum quality of service standards, e.g. regarding waiting times of customer services, times to delete private data, fees, how to contact the data management center's service, whom to contact in case of complaints.

Compliance to the protocol would allow companies to show a "Privacy-Safe Badge" on their Web site, which would immediately be recognised by surfers (see [60, 61]). Showing the privacy badge would probably become fundamental for certain categories of companies operating the Net (e.g. search engines, social networks, banks, etc.). Not having the badge, could make a relevant difference in the trust level of customers. Moreover, it is easily foreseeable, as data-mining activities become more pervasive in the future, the importance of the badge would eventually extend to other general purpose Web sites.

The badge would be granted by newly created (ideally super-national) rating institutions, which should also have the authority to enforce the standards related to the respective security badge by inspection. Such an institution will be the only legal parties empowered to issue privacy badges and revoke them in case of misconduct. In future, the collection of personal data on the Internet without a proper badge could be considered an illicit activity and insofar be sanctioned by users accordingly.

To obtain the badge, interested parties would have to demonstrate that they possess both, the *ethical* and *technical* standards necessary to accomplish such a delicate data mining task. After proper checking, and depending on the purpose of the data collection specified in the harvesting declaration, different types of badges could be issued. Each badge would also be linked to a standardized user licence.

In order to add dynamism and a more democratic taste, the badge could foresee user ratings and comments. These opinion feedbacks per se, would not generate legal consequences for the owner of the badge, but would help to detect misconducts earlier and to alarm the community, and it would trigger inspection procedures by the issuer of the badge.

Finally, users should be able to a-priori set their preferences and conditions on their browsing devices under which participating or not to data-collection campaigns[3]. Browsers would immediately examine the badge of any visited Web site, comparing it with its stored preferences and automatically notify any threat to the privacy of the users. Besides, this would solve the notorious issues of unreadable or over-technical "Term of Services" conditions, which should not any longer be read directly by users themselves.

3.6 Privacy-respecting social networks

Social networking has been rapidly spreading in the past years despite frequent warnings regarding a lack in privacy protection. Recently, for example, somebody succeeded with downloading 100 Million user profiles and upload the dataset for free download by everybody [111]. It is often claimed that users simply do not seem to care about uploading private information to the Internet. However,

- this does by far not apply to everybody (in fact, most computer users still do not have social networking profiles),
- the terms of use have changed since most of them have uploaded their private data (e.g. photographs),
- some users may assess the comfort of the service provided by social networking sites higher than the current side effects, but this may change over time.

Besides, a recent empirical study has impressively demonstrated that people *do* care about the use of their activity data [84].

It certainly appears necessary to have alternative technical solutions for social networking, which protect privacy better. A first project of this kind is DIASPORA [85], which wants to decentralize the storage of sensitive information.

Privacy-protecting social networks could be imagined as follows: Individuals would only see part of the network. Individuals and communities could determine what can be seen to outsiders of the community and to whom (friends, friends of friends, second-next-nearest neighbors, or everyone; same with business partners). Depending on this, certain kinds of information would not be visible to outsiders, others would be (as communities may want to gain new members). In essence, surfing in social networks would be like travelling between communities, and this would feel like visiting other countries. While certain things would be visible, others (the private part of the information) would remain hidden to strangers (as private houses are).

3.7 Summary

In essence, many of the problems of the Internet today result from Web2.0 and other applications, which the Internet was originally not designed for. Consequently, current technical solutions are insufficient. A new way of organizing the Internet appears to be needed and possible. Suitable solutions can be developed by transferring concepts of social self-organization to the design of the future Internet. This constitutes an interesting challenge within the research field of techno-social systems.

[3]For example, willingness to allow collection of personal data only for scientific purposes.

4 Recommended legal regulations

Currently, data about people are probably processed, used and misused in any conceivable way. Since regulations are insufficient and heterogeneous, the situation has sometimes been paraphrased as Wild Wild Web. It is therefore not surprising that the EU Fundamental Rights and Citizenship Commissioner Viviane Reding has recently pointed out [9] that Europe needs more harmonization regarding a data protection law. Determining the best routes towards this goal deserves targeted research. However, as the problems are actute, action needs to be taken soon. Therefore, the following sections make a number of suggestions.

Given the problems of the current Internet and the foreseeable future developments, data collection for research or for business should be regulated taking into account privacy, legal requirements, science's and business' interests. We foresee that methods of data collection should be open, controllable and verifiable by legal authorities and the public. Legal procedures and the law should establish what type of data can be collected, what type of data may not be collected, and how the sensitive part of the collected data must be hidden from people or organizations collecting them at each point of the data collection procedure. Methods of warranting the safety of sensitive data should be public and should be verifiable at all times before, during, and after collection. For example, we recommend to work out legal regulations for the following:

- Data storage, access, processing and usage standards should be fixed for public, commercial and private entities operating in a certain country. Transparency regarding the storage, access, processing and use of data should be enforced. In particular, there should be a binding public declaration of what kind of private data are being stored, processed and used, and how this is done.

- Personal data should always be stored in an encrypted way. However, it should be made easy to inform oneself free of charge about the data determined and stored by other individuals, companies or institutions, and how these data are accessed, protected and used. Therefore, technical solutions should be required, which allow individuals to access (and decrypt) their personal data on-line and to delete data one does not want to be stored (if there is not a law that requires such storage). Further on, it should be easy to opt out from the determination, storage and/or processing of certain kinds of data. It should be possible to sanction violations of this right efficiently, and affected individuals should be properly compensated.

- One should establish standards ensuring informed consent of users with the data an information system is determining, storing, or processing. Users should not be forced to agree with a storage, processing and use of information that is not technically required for the services a user wants to get. For example, providers of media contents should not force customers to reveal their identities (and effectively their preferences via the contents they buy), as the contents or services can also be paid for anonymously. Putting it differently, companies should be required to offer, in a clearly marked way, options to customers that allow them to choose at any time between a data-rich variant (providing the service provider with many detailed individual data) and a data-poor (privacy-protecting) variant without artificially created disadvantages (which would effectively force customers to reveal their data). Within fair limits, it would be acceptable though to charge a higher price for data-rich services to users, who have decided for the data-poor variant themselves.

- Licence and usage agreements of software products and information services should be regulated and standardized. As most users do not read or understand the terms of use, and as they do not have any chance to negotiate these conditions, there should be a few (certainly less than ten) standard licences, which should be signalized by a color or other codes, whenever a certain software or information-based service is used. Alternatively, softwares and services should be rated by independent agencies based on the benefits users can expect from them and the degree to which privacy and confidentiality are potentially affected.

- It would be useful to define the individual and corporate responsibilities for damages created in the virtual or real world by activities in the Internet.
- However, considering the fact that the content of a file is revealed only when it is accessed, it should not be possible to punish people for the access of contents, if the contents are not warned of in advance in a sufficient and qualified way (e.g. based on the rating and reputation system suggested above). In other words, users should be protected from legal traps in the Internet.
- Considering the abundance of free contents in the Internet, it is advised to implement a copyright, which considers the facts of modern information systems and requires copyright holders to make proper attempts to protect their products from unauthorized access (e.g. to indicate their copyrights, encode electronic files, and offer simple, fair, and anonymous payment procedures).
- Compensations for privacy violations would have to be fixed, and legal procedures would need to be simple and effective to allow people to protect their rights. For example, fines to companies, which sell private data without authorization, should be significantly higher than the likely profit they can make on such business.
- The priorities in cases of conflicts of interest should be worked out clearly (protection of individual human rights comes before collective public interests, which comes before institutional interests of companies or political parties, which comes before individual interests).
- Legal regulations should protect individuals against discrimination based on private data and guarantee an efficient compensation in case of violations.
- The introduction of class action would allow users to better defend their rights at court against individuals, companies, or institutions violating them, but the implementation should consider that the way attorneys are compensated and the way discovery is organized in civil procedures largely determines how desirable and effective class action lawsuits are.
- Unique virtual identities/electronic signatures should be offered for everybody.
- It should be required to specially mark Web links that are redirected to contents with a different character, or Internet services that are changing their character (e.g. from non-commercial to commercial), or the use of pseudonyms that have been used before by others.
- It may be useful to fix a statute of limitations, i.e. a time period after which violations of Internet-related regulations cannot anymore be sued. These time periods should increase with the seriousness of the violation and its consequences. It should also depend on whether the effect of the violation was in the past or relevant for the presence and future as well.
- There should be legal procedures regarding the random and targeted control of the fulfilment of legal standards regarding the storage, access, processing and use of private data.
- Conditions should be worked out for imposing access restrictions or forced decryptions of suspicious Internet contents, in case there is evidence that they seriously threaten the public security (such as instructions how to build bombs). Such measures, their extent and results would have to be reported to the public, and individuals would have to be compensated, if it turned out that they were unjustified.
- Companies receiving public money should be required to make data of public interest available for research, after they have been processed in a way that removes sensitive information (see the above sections on how this can be done).

- It would be good to have neutral, publicly controlled third-party infrastructures, which allow to perform anonymous Web experiments and data mining.
- Special procedures should be defined for cases, where access to original or sensitive anonymized data is justified and required (e.g. for certain kinds of research of public interest). A good example is the way in which Harvard University regulates the access and processing of the data of the Framingham Heart Study, which allowed scientists to discover social processes promoting the spreading of obesity, smoking, depression, or happiness, to mention just a few relevant examples of gained insights that can be beneficial for individuals and the public [115].
- There should be a fair right of information and participation in social activities mediated by ICT systems. For this reason, information businesses directed at a mass audience and with a large market share should be required not to discriminate and exclude certain user groups through inappropriate pricing schemes or terms of use (e.g. the requirement to agree with the arbitrary use or transfer of personal data or the requirement to allow for cookies, where this is technically not needed to provide the requested service). Individuals should always have the possibility to opt out of data uses they do not agree with, without losing access to information services not requiring these data.

5 Recommended infrastructures and institutions

In order to have a powerful, largely self-regulating Internet, the following kinds of institutions would be useful to have:

1. Public data centers, which perform a neutral and independent data collection that is not driven by the need to make money, but serves the purpose to inform the public in the best possible way. Such a system could implement the reputation, community formation, sanctioning and privacy respecting mechanisms discussed before in connection with the concept of a self-organizing Internet.
2. Research centers, which study what can be done with publicly available data, to assess the potentials and risks. These centers should also develop the technology of the self-organizing Internet sketched above.
3. Publicly controlled, neutral institutions, which can serve as independent third parties in experimental designs that ensure anonymity (see Sec. 2.5).
4. Independent quality audit centers, which evaluate the level to which companies protect privacy and provide good services and fair terms of use.
5. One or several complaint center(s), which collects complaints of Internet users and can take action against illegal or unethical practices. These centers should be well connected with the public media.
6. An ethical committee, which assesses risks of information technologies and markets. It should set ethical standards regarding the storage and processing of data and support the preparation of required legal regulations.
7. A center working out contingency plans for the case of large-scale failures of information and communication infrastructures, e.g. due to denial of service attacks, spam, viruses, trojan horses, worm or phishing problems, or solar-storm-related failures of electronic systems.
8. A committee working out suggestions for legal settings, as the need for institutional regulations arises through new technological developments.

6 Summary

Socio-economic data mining has a great potential in terms of gaining a better understanding of problems that our economy and society are facing, such as financial instability, shortages of resources, or conflicts. Without large-scale data mining, progress in these areas seems hard or impossible. Therefore, a suitable, distributed data mining infrastructure and research center should be built in Europe.

Reality mining provides the chance to adapt more quickly and more accurately to changing situations. For example, it will facilitate a real-time management of challenges like evacuation scenarios or economic stimulus programs. Further opportunities arise by individually customized services, which however should be provided in a privacy-respecting way. This requires the development of novel ICT (such as a self-organizing Internet), but most likely new legal regulations and suitable institutions as well.

As long as such regulations are lacking on a world-wide scale (and potentially even thereafter), it is in the public interest that scientists explore what can be done (in a positive and negative sense) with the huge data available about virtually everybody and everything. Big data do have the potential to change or even threaten democratic societies. The same applies to sudden and large-scale failures of ICT systems. Therefore, dealing with data must be done with a large degree of responsibility and care. Self-interests of individuals, companies or institutions have limits, where the public interest is affected, and public interest is not a sufficient justification to violate human rights of individuals. Privacy is a high good, as confidentiality is, and damaging it would have serious side effects for society.

Acknowledgements

The authors of this White Paper are grateful to Karl Aberer, Andras Lörincz, Panos Argyrakis, Endre Bangerter, Andrea Bassi, Stefan Bechtold, Bernd Carsten Stahl, Rui Carvalho, Markus Christen, Mario J. Gaspar da Silva, Fosca Giannotti, Aki-Hiro Sato, David-Olivier Jaquet-Chiffelle, Daniel Roggen, Themis Palpanas, Elia Palme, Jürgen Scheffran, David Sumpter and Peter Wagner.

References

1. L. Backstrom, C. Dwork and J. Kleinberg, Wherefore Art Thou R3579X? Anonymized Social Networks, Hidden Patterns, and Structural Steganography. *Proc. 16th Int. World Wide Web Conference* (2007)

2. Apple confirms $1bn data center.
   http://www.theregister.co.uk/2009/06/04/apple_1bn_north_carolina_data_center/

3. NSA plans massive, 65MW, $2bn data center in Utah.
   http://www.theregister.co.uk/2009/07/03/new_nsa_data_center/

4. Microsoft consumes Chicago data center.
   http://www.theregister.co.uk/2009/05/20/ascent_ch2_datacenter/

5. Google admits Scandinavian data center landing.
   http://www.theregister.co.uk/2009/03/05/google_finland_data_center/

6. Google pays $51.7m for newspaper destruction metaphor.
   http://www.theregister.co.uk/2009/02/12/google_buys_defunct_paper_mill/

7. Intel sees future in Mega Data Center.
   http://www.theregister.co.uk/2009/02/18/the_intel_cloud/

8. D. Helbing and S. Balietti, From social simulation to integrative system design. Visioneer White Paper (2010), see http://www.visioneer.ethz.ch.

9. EU Commission plans more harmonisation of data protection law. http://www.out-law.com/default.aspx?page=11228

10. P. Bajaria and J. Yeo , Auction design and tacit collusion in FCC spectrum auctions. *Information Economics and Policy* 21:**2** 90-100 (2009)

11. C. Schultz, Transparency and tacit collusion (2001)

12. B. Kluger and S.B. Wyatt, Preferencing, Internalization of Order Flow, and Tacit Collusion: Evidence from Experiments. *Journal of Financial and Quantitative Analysis* 37:**3** pg:449 (2002)

13. M. Michael, J.E. Moreira, D. Shiloach and R.W. Wisniewski, Scale-up x Scale-out: A Case Study using Nutch/Lucene. *Parallel and Distributed Processing Symposium, IEEE International* (2007)

14. L. A. Barroso and Hölzle, The Datacenter as a Computer: An Introduction to the Design of Warehouse-Scale Machines. *Morgan & Claypool Publishers* (2009)

15. A. Jacobs, The pathologies of big data. *ACM Queue*, 7:**(6)** (2009)

16. SSD Myths and Legends - "write endurance". http://www.storagesearch.com/ssdmyths-endurance.html.

17. A. Mazlouimian, D. Helbing, Y-H. Eom, S. Lozano and S. Fortunato, How citation boosts trigger scientific paradigm shifts. (in preparation) (2010).

18. D. Helbing, M. Treiber, and N. J. Saam, Analytical investigation of innovation dynamics considering stochasticity in the evaluation of fitness. *Physical Review E* **71**, 067101 (2005).

19. J. Lorenz, H. Rauhut, F. Schweitzer, and D. Helbing, How social influence undermines the wisdom of crowds. Submitted (2010).

20. S. E. Asch, Studies of independence and conformity: a minority of one against a unanimous majority. *Psychological Monographs*, 70:**9** (1956).

21. F. Winter, H. Rauhut and D. Helbing, How norms can generate conflict. *Jena Economic Research Papers* (2009).

22. D. Helbing and A. Johansson, Cooperation, norms, and conflict: A unified approach. *SFI Working Paper* #09-09-040 (2009).

23. D. Helbing and W. Yu, The outbreak of cooperation among success-driven individuals under noisy conditions. *Proceedings of the National Academy of Sciences USA (PNAS)* 106**(8)**, 3680-3685 (2009).

24. D. Helbing, W. Yu and H. Rauhut, Self-organization and emergence in social systems. Modeling the coevolution of social environments and cooperative behavior. *SFI Working Paper #09-07-026* (2009).

25. Four million British identities are up for sale on the Internet. http://technology.timesonline.co.uk/tol/news/tech_and_web/the_web/article6718560.ece

26. E. Diener and R. Crandall, Ethics in social and behavioral research. *University of Chicago Press, Chicago* (1978).

27. Geambasu R., Kohno T., Levy A. and Levy H. M., Vanish: Increasing Data Privacy with Self-Destructing Data. *Proceedings of the USENIX Security Symposium*, Montreal, Canada, August (2009)

28. It seems it not so easy to clear ones name on-line, even when trying hard. This is specially true for traces left on social network Web sites, for which specific applications, such as http://suicidemachine.org/, have been created in order to accomplish this task. For a discussion based on a true story see http://ask.slashdot.org/story/09/12/10/2115238/Best-Way-To-Clear-Your-Name-Online.

29. R. Axelrod, *The Evolution of Cooperation*, Basic Books, 1984 pp. 169–170.

30. Data, data everywhere, *The Economist*, Feb 25th 2010

31. L. Lessig, Against Transparency, *The New Republic*, 9th Oct, 2009. http://www.tnr.com/article/books-and-arts/against-transparency

32. Symantec Internet Security Threat Report. http://www.symantec.com/business/theme.jsp?themeid=threatreport.

33. A. Janc and L. Olejnik , Feasibility and Real-World Implications of Web Browser History Detection

34. G. Wondracek, T. Holz, E. Kirda and C Kruegel, A Practical Attack to De-Anonymize Social Network Users. Technical Report TR-iSecLab-0110-001

35. Google admits it accidentally gathered WiFi data. http://www.ft.com/cms/s/2/8a23b394-5fab-11df-a670-00144feab49a.html

36. Google to hand over intercepted data. http://www.ft.com/cms/s/2/db664044-6f43-11df-9f43-00144feabdc0.html

37. Lawyers Claim Google Wi-Fi Sniffing "Is Not an Accident". http://gizmodo.com/5554960/lawyers-claim-google-wi+fi-sniffing-is-not-an-accident

38. Wi-Fi Data Captured By Google Street View Cars Included Passwords. http://gizmodo.com/5567460/wi+fi-data-captured-by-google-street-view-cars-included-passwords

39. Ethical Guidelines, Social Research Association", (2003). http://www.the-sra.org.uk/ethical.htm

40. Dench Sally, Iphofen Ron, Huws Ursula, An EU Code of Ethics for Socio-Economic Research, *The Institute of Employment Studies*, 2004

41. The British Psychological Society, Report of the Working Party on Conducting Research on the Internet, 2007. http://www.bps.org.uk/the-society/code-of-conduct/

42. Senator calls on FTC to tackle social-net privacy. http://news.cnet.com/8301-13577_3-20003415-36.html

43. The Electronic Frontier Foundation published a timeline of Facebook's privacy policy modifications over the years. http://www.eff.org/deeplinks/2010/04/facebook-timeline

44. Watchdog files complaint over Facebook 'privacy' settings.
http://www.theregister.co.uk/2009/12/17/epic_facebook_privacy_complain/

45. Exclusive: Google, CIA Invest in 'Future' of Web Monitoring.
http://www.wired.com/dangerroom/2010/07/exclusive-google-cia/

46. X. Su and T. M. Khoshgoftaar, A Survey of Collaborative Filtering Techniques, *Advances in Artificial Intelligence* (2009).

47. E. J. Candes and T. Tao, The power of convex relaxation: Near-optimal matrix completion. *IEEE Trans. Inform. Theory*, 56: 2053-2080 (2009).

48. Privacy and identity management for life. Eds. M. Bezzi et al. Springer (2009).

49. G. Ziegler, C. Farkas, and A. Lörincz, A framework for anonymous but accountable self-organizing communities, Information and Software Technology, 48: 726-744 (2006).

50. Apple's Worst Security Breach: 114,000 iPad Owners Exposed.
http://gawker.com/5559346/apples-worst-security-breach-114000-ipad-owners-exposed

51. T-Mobile confirms biggest phone customer data breach.
http://www.guardian.co.uk/uk/2009/nov/17/t-mobile-phone-data-privacy

52. EU warns on Facebook privacy.
http://www.nytimes.com/2009/01/27/technology/27iht-facebook.4-417144.html

53. German minister warns Facebook over privacy rules.
http://blog.foreignpolicy.com/posts/2010/04/05/german_minister_warns_facebook_over_privacy_rules

54. Cattuto C., Van den Broeck W., Barrat A., Colizza V., Pinton J.-F. and Vespignani A., Dynamics of person-to-person interactions from distributed RFID sensor networks. *PLoS ONE* 5**(7)**: e11596 (2010)

55. J. Krumm, A survey of computational location privacy. Personal and Ubiquitous Computing 13 (6), 391–399 (2009).

56. Internet Reputation Management: neutralize negative publicity.
http://www.internet-reputation-management.com/

57. Reputation Management Consultants.
http://www.reputationmanagementconsultants.com/

58. Reputation Defender. http://www.reputationdefender.com/

59. Squidoo: Internet Reputation Management.
http://www.squidoo.com/internet-reputation-management

60. European privacy seals for IT products and IT-based services.
https://www.european-privacy-seal.eu/

61. Ixquick: the world's most private search engine. http://ixquick.com/

62. RottenNeighbor.com was a website created to post information about neighbors and find information about new potential neighbors before moving. Launched in July 2007, it was discontinued in July 2009.

63. Digging into Data. http://www.diggingintodata.org/
64. Transparency is at the heart of this Government. Data.gov.uk is home to national & local data for free re-use. http://data.gov.uk.
65. Data.Gov Empowering people. http://www.data.gov.
66. Dataverse Project: An Open-Source Application for Publishing, Citing and Discovering Research Data. http://thedata.org/.
67. Apache WSIF: Web Service Invocation Framework. http://ws.apache.org/wsif/.
68. ETH Financial Crisis Observatory. http://www.er.ethz.ch/fco/index.
69. 123 People. http://www.123people.com.
70. Oakland Crimespotting is an interactive map of crimes in Oakland and a tool for understanding crime in cities. http://oakland.crimespotting.org.
71. Government requests directed to Google and YouTube. http://www.google.com/governmentrequests/.
72. Mechanical Turk is a market place for work. https://www.mturk.com/mturk/welcome.
73. Micro Payment: professional payment provider. http://micropayment.de.
74. Vanish: self-destructing digital data. http://vanish.cs.washington.edu.
75. Did you watch porn? http://www.didyouwatchporn.com.
76. What the Internet knows about you. This page checks your browser history and determines which of the 5000 most popular Internet websites you've recently visited. http://www.whattheinternetknowsaboutyou.com.
77. Peter Eckersley, How Unique Is Your Web Browser? *Electronic Frontier Foundation* (2009)
78. Panopticlick: How unique and trackable is your browser? https://panopticlick.eff.org.
79. EFF: Forget cookies, your browser has fingerprints. http://www.computerworld.com/s/article/9176904/EFF_Forget_cookies_your_browser_has_fingerprints
80. WUALA, Backup. Store. Share. Access Everywhere. http://www.wuala.com/.
81. Tor: anonymity online. http://www.torproject.org/.
82. Freenet, the free network. http://freenetproject.org/.
83. QLectives (Quality Collectives). http://www.qlectives.eu.
84. Datenschutz fr iPhone-Apps. http://www.ethlife.ethz.ch/archive_articles/100930_MBusiness_Apps_sch/index
85. DIASPORA, The privacy aware, personally controlled, do-it-all, open source social network. http://www.joindiaspora.com/
86. Security Focus. http://www.securityfocus.com
87. ICKN Galaxy Advisors. http://ickn.org

88. PostRank: Intelligence from the social web. http://www.postrank.com/

89. Soziale Netzwerke verraten künftiges Käuferverhalten. http://www.tagesanzeiger.ch/digital/internet/Soziale-Netzwerke-verraten-kuenftiges-Kaeuferverhalten/story/19928880

90. M.M. Gaber, A. Zaslavsky and S. Krishnaswamy, Mining data streams: a review. *ACM SIGMOD Record archive*, 34:**2** 18–26, (2005).

91. A. Bifet and R.K. August, Data Stream Mining: A Practical Approach. *The University of Waikato* (2009).

92. J. Leskovec, L. Backstrom and J. Kleinberg, Meme-tracking and the dynamics of the news cycle. *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 497–506, (2009).

93. R. Agrawal and R. Srikant, Privacy-preserving data mining. *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pp. 439450 (2000).

94. P. Samarati and L. Sweeney, Generalizing Data to Provide Anonymity when Disclosing Information. *Proceedings of the Seventeenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*. ACM Press (1998).

95. C. C. Aggarwal and Philip S. Yu, Privacy-Preserving Data Mining: Models and Algorithms. *Springer*, pp. 530, (2008).

96. M. Atzori, F. Bonchi, F. Giannotti and D. Pedreschi, Anonymity preserving pattern discovery. *VLDB Journal*, 17:**4**, pp. 703–727, (2006).

97. B.C.Chen, D. Kifer, K. LeFevre and A. Machanavajjhala, Privacy-Preserving Data Publishing. *Foundations and Trends in Databases* Vol. 2, Nos. 12 pp.1167 (2009)

98. B.-C. Chen, D. Kifer, K. LeFevre and A. Machanavajjhala, Privacy-Preserving Data Publishing (Survey).*Foundations and Trends in Databases* 2, Nos. 12 1167 (2009).

99. A. Narayanan and V. Shmatikov, Robust De-anonymization of Large Sparse Datasets. *IEEE Symposium on In Security and Privacy SP 2008*. IEEE Symposium. pp. 111-125. (2008).

100. R.Jones, R. Kumar, B. Pang, A.Tomkins, Vanity fair: privacy in querylog bundles. *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management* pp. 853-862 (2008).

101. F. Giannotti and D. Pedreschi, Mobility, Data Mining and Privacy: Geographic Knowledge Discovery. *Springer* pp. 410 (2008).

102. A. Monreale, G. Andrienko, N.Andrienko, F. Giannotti, D. Pedreschi, S. Rinzivillo and S. Wrobel, Movement Data Anonymity through Generalization. *Transactions on Data Privacy* 3:**2** pp. 91–121 (2010). http://www.tdp.cat/issues/abs.a045a10.php

103. Project Gaydar. http://www.boston.com/bostonglobe/ideas/articles/2009/09/20/project_gaydar_an_mit_experiment_raises_new_questions_about_online_privacy/.

104. Privatsphäre als Luxusgut. http://www.nzz.ch/blogs/nzz_blogs/betablog/privatsphaere_als_luxusgut_1.7266824.html

105. D. Helbing et al., Dynamic decision behavior and optimal guidance through information services: Models and experiments. Pages 47-95 in: M. Schreckenberg and R. Selten (eds.) Human Behaviour and Traffic Networks *Springer, Berlin* (2004).

106. D. Helbing and M. Christen (2010). Mit Rauschen und Reibung gegen finanzielle Blasen, submitted to Wirtschaftswoche.

107. M. Mäs, A. Flache, and D. Helbing (2010) Individualization as driving force of clustering phenomena in humans. PLoS Computational Biology, in print.

108. C. Frankfort-Nachmias and D. Nachmias, "Research Methods in the Social Sciences" (Worth Publishers, New York, 2008), Chap. 4: "Ethics in Social Research".

109. Statement of Ethical Practice for the British Sociological Association, BSA, the British Sociological Society, 2002. http://www.britsoc.co.uk/equality/Statement+Ethical+Practice.htm

110. Google chief: Only miscreants worry about net privacy. http://www.theregister.co.uk/2009/12/07/schmidt_on_privacy/

111. Details of 100m Facebook users collected and published. http://www.bbc.co.uk/news/technology-10796584

112. Recommendations of the Association of Internet Researchers (AoIR) Ethics Working Committee. http://www.aoir.org/reports/ethics.pdf

113. D. Helbing and W. Yu, The future of social experimenting. *PNAS* 107**(12)** 5265-5266, (2010).

114. The Future of Social Experimenting: The Full Story. http://www.soms.ethz.ch/research/socialexperimenting

115. K.P. Smith and N.A. Christakis, 'Social Networks and Health, *Annual Review of Sociology* 34: 405-429 (2008).

116. D. Helbing, W. Yu, K.-D. Opp and H. Rauhut, The emergence of homogeneous norms in heterogeneous populations. *American Journal of Sociology*, submitted (2010).