Mining Data Related to Children: Ethical Challenges

Daya C. Wimalasuriya
Department of Computer Science and Engineering,
University of Moratuwa, Sri Lanka
chinthana@cse.mrt.ac.lk

Dejing Dou
Department of Computer and Information Science,
University of Oregon, USA
dou@cs.uoregon.edu

Dilini Wimalasuriya
School of Management,
Northshore College of Business and Technology, Sri Lanka
dilini@northshore.lk

ABSTRACT

Data mining on children's computer usage is increasingly being highlighted as an important ethical concern by mass media. Not surprisingly, such commentaries are generally negative in tone and point out the questionable practices employed by organizations engaged in this exercise. On the other hand, the data mining community often highlights the benefits that can be derived from this. In this paper, we review the literature on this area to identify the key ethical concerns raised by various stakeholders and provide recommendations for what the data mining community can do to address them.

Categories and Subject Descriptors

K.4.1 [**Public Policy Issues**]

General Terms

Human Factors, Legal Aspects.

Keywords

Ethics, Data Mining, Children

1. INTRODUCTION

In mid May 2014, the online news and opinion portal Politico published a widely cited article titled "Data mining your children" [1]. Written by Stephanie Simon, the article raises several ethical questions regarding the data mining practices of various education software developers. It analyzes the data mining practices of organizations ranging from relatively new startups such as LearnSprout, which stores and analyzes student information such as attendance records to the well-established institutions such as the Khan Academy. "Data is the real asset" – the article quotes Sal Khan, founder of the Khan Academy, who is said to have made this statement in an academic conference in the fall of 2013. On the Khan Academy, the article states "It's free. But users do pay a price: In effect, they trade their data for the tutoring". This statement sums up the tone and the message of the entire article.

What the Politico article shows are the perils encountered by the organizations that engage in mining data related to children. The Khan Academy, which is often hailed for revolutionizing education technology, is easily cast on a negative light because of the manner in which it handles the vast amount of data it collects on children. Because the implication is that children - who are vulnerable than adults and whom the society is obliged to protect - are exploited, the damage caused to the credibility of a data mining organization because of such a disclosure is significantly higher than by a disclosure regarding adults.

The Politico article was not the only publication that has raised ethical concerns about mining data related to children in the recent past. Just about two months before the publication of the article, it was widely reported in new media that Google has been sued for scanning student emails under Google Apps for Education to build up profiles of the users [2]. The legal argument has been that the processed information may contain students' academic records which are protected by the Family Educational Rights and the Privacy Act (FERPA) but the underlying ethical argument is that mining data collected from children in an educational setting is wrong. In response, Google has decided to stop the practice of scanning student emails altogether when they are using Google Apps for Education. It should be noted that in this case, most students have not even seen personalized advertisements as the default option has been to turn off ads but that has not stopped the grievances from arising.

Moreover, such resentments towards mining children's data for marketing purposes arise not in a vacuum but in a wider context of ethically questionable marketing practices towards children. For example, according to the CEO of prism communications "they aren't children so much as what I like to call 'evolving customers'" [3]. Such statements are often vilified in the popular literature and builds up a caricature of a cold-hearted marketing industry that exploits children. When concerns are raised about the data mining practices on children, whose outputs are widely used for marketing purposes, it is not difficult for the public to link such practices with the negative aspects of the marketing industry and view the data mining community and the industry on the same negative light. As mentioned earlier, since the wronged party is children, who ought to be protected by the society, this often invokes an emotional response. This shows why the data mining community and the industry has to be very careful in dealing with children's data.

On the flip side, data mining on children can provide benefits to children and the wider society. The data mining practices of the Khan Academy have been brought into question, but its methods - often supported by data mining – have helped thousands of students. The same can be said about the Google Apps for Education. Therefore, what is needed here is to address the legitimate ethical concerns related to mining data collected from children while not hampering the benefits that can be derived. In

KDD 2014 Data Ethics Workshop, August 24, 2014, New York, NY, USA.

this paper, we attempt to help the data mining community in this exercise.

It is important to point out that by "data related to children", we mean not only the personal data collected by websites from children such as name, age, address, height, weight, etc., but also any data collected by the website from children through its interaction with the website. For example, in a social network this includes friend lists, posts, etc. In an educational website, data such as the time spent on a particular lesson are included. This often results in a huge data set in a given setting. We look at the ethical challenges associated with mining such data sets.

The rest of the paper is organized as follows. In Section 2, we delve into the background in legal, marketing and technical aspects related to mining children's data. In Section 3, we summarize the ethical concerns raised and suggest strategies that can be adopted by the data mining community to counter these. Some of these are based on the literature and industry practice while others are our own recommendations. In Section 4, we provide some concluding remarks on the topic.

2. BACKGROUND

2.1 Laws

Children's Online Privacy Protection Act of 1998 (COPPA) is the primary legislation in the United States regarding the ethics of dealing with children using the Internet. By definition, it applies to websites under US jurisdiction and websites in other countries that are directed to children in the US or knowingly collect information from children of the US. However, it has been observed that even foreign companies that are not subject to the law are complying with it. As such, this law is a good indicator of laws on children's privacy on the Internet across the world.

The law applies to children under the age of 13. The general idea of the law is that while websites can collect personal information of children under the age of 13, they have to obtain the consent of the parents or the guardians and have to be extra careful in dealing with the information collected. The following are the main requirements the law imposes on websites dealing with children under the age of 13 [4].

- Provide a clear and comprehensive online privacy policy on dealing with personal information of children under the age of 13.
- Make reasonable efforts to provide direct notice to parents of the operator's practices with regards to children's information.
- Obtain verifiable parental consent of children under the age of 13.
- Provide reasonable means for the parents to review the child's information and refuse to permit its future use.
- Establish and maintain reasonable procedures to protect the confidentiality, security and integrity of the personal information collected from children.
- Retain personal information of collected online information from children only as long as necessary for the purpose it was collected for and delete it using reasonable measures after that to protect against unauthorized access.
- Not conditioning a child's participation on an online activity based on the child providing more information than is reasonably necessary to participate in that activity.

When studying the requirements of the Act, it is quite clear that complying with them requires a lot of work although they make a lot of common sense. For instance, requiring the companies to delete the information once they are not necessary for the purpose they were collected for appears reasonable but technically, determining when the information becomes unnecessary is difficult. Even permanently deleting online information is not straight-forward or easy. Other requirements of the Act, such as obtaining parental consent, raise similar difficulties. Since these are legal requirements, not complying with them makes organizations liable for legal action.

The primary result of the Act has been that most website operators, including major players such as Google and Facebook, completely disallow the use of their services by children under the age of 13. Complying with the Act is seen as extremely difficult and avoiding dealing with personal data of children under the age of 13 is seen as the better option.

COPPA can be seen as a warning to the data mining community and the industry on why they have to be careful in mining children's data: if the public opinion turns strongly against the current data mining practices on data taken from children between the ages of 13 and 18, that might result in a legislation similar to COPPA for those children as well. While the framers of the legislation will try not to unreasonably prohibit web sites from interacting with these children, the experience of COPPA shows that the effect of such a law may exactly be this.

The Family Education Rights and Privacy Act of 1974 (FERPA) is another legislation that is relevant to children's data. It applies to students' educational records held by an institute. The general idea of the law is that these records can only be released with the consent from the student (if she is over 18) or from the parents. It allows some exceptions on occasions such as a request from a school a student is transferring to.

FERPA is not related to mining children's data to the extent of COPPA, but the lawsuit against Google on scanning students emails in Google Apps for Education (mentioned in Section 1) shows that it can be relevant. In addition, the Politico article described above, questions the practices of Learn Boost, which allows teachers to upload student education records to its servers and then share them with parents and other parties in the context of these information being covered by FERPA.

2.2 Marketing for Children

Walt Disney and Ray Kroc (founder of the McDonald's chain) are often mentioned as the pioneers in the field of marketing for children. It has emerged as a concept of marketing in the 1980's [5]. Several books have been written on the subject.

According to McNeal, the following are the main stages of a child evolving as a customer [6].

- From age 1: Accompanying parents and observing.
- From age 2: Accompanying parents and requesting.
- From age 3: Accompanying parents and selecting with permission
- From age 4: Accompanying parents and making independent purchasing decisions.
- From age 5: Going to store and making independent purchases.

Beder has identifies the following three types of purchasing decision as being affected by marketing for children [7].

- Purchases made by children by themselves (e.g., sweets, toys, fast food, clothes, shoes)
- Purchase decisions of parents affected by children (e.g., furniture, electrical appliances, vacation, vehicles)

- Purchases made by children once they grow up. (e.g., vehicles)

She states that each of these categories account for billions of dollars worth annual sales. She further states that the third category as being particularly important for marketers as it allows them to build brand loyal customers from childhood who are shown to be more beneficial than customers converted from competitors at adulthood.

While these facts show the opportunities associated with marketing for children, it has to be recognized that a lot of people find marketing for children, when their minds are still developing and hence vulnerable, objectionable. Based on such concerns, Beder argues that marketing for children should be carefully restricted. In particular, she argues that marketing for children under the age of 9, including on the television and on the Internet, should be banned.

Because of COPPA, it can he assumed that web sites do not keep profiles of children under the age of 13. Therefore, ethical issues regarding marketing for very young children should not arise in the context of mining children's data. However, it is important to recognize the concerns on whether it is socially acceptable to use sophisticated marketing techniques on children. While the critics are likely to be more tolerant of the practice for older children (for example, Beder has not called for banning advertisements for children above the age of 9), it is important to recognize that concerns linger. This is useful in coming up with strategies on how to deal with data of children older than 13.

The task force appointed by the American Psychological Association (APA) to study advertising on children and its recommendations are of special interest in investigating the ethics of mining children's data. The task force had been appointed in 2000 after ethical concerns were raised about psychologists working with advertisers to fine tune the marketing strategies aimed at children [8]. The task force in its report published in 2004 [9] has made several recommendations for public policy including restricting advertising primarily directed towards audiences of young children (children aged 8 and younger) and requiring advertising disclaimers in language that can be comprehended by children. It encourages psychological research to further examine the effects of marketing on children including influences of the new interactive media environments (the Internet). On the issue of psychologists working with advertisers, the report recommends that the APA "undertake efforts to help psychologists weigh the potential ethical challenges involved in professional efforts to more effectively advertise to children, particularly those children who are too young to comprehend the persuasive intent of television commercials".

On a follow up analysis of the report, Kramer states that APA cannot have ethical standards that restrict the ability of its members to use their skills to earn a living because of antitrust laws [10]. Therefore, she states that the question is how APA members can use principles of ethics such as "psychologists must strive to benefit those with whom they work and take care to do no harm" if they choose to work with advertisers.

2.3 Technology

Personalized advertisements are one of the most visible applications based on user data mining. The general idea here is building up a user profile based on the information available to a website (e.g., words of emails, products viewed, age information) and serving advertisements that are most likely to be of interest to that user. This is beneficial to the users as well as the advertisers. Depending on the comprehensiveness of the user profile and range of advertisers available, advertisements can be micro-targeted to different levels of granularity.

HTTP Cookies are normally used in serving personalized advertisements. They convey information such as what ad campaigns the user was exposed to previously. While no personal information are stored in the cookie files, they do enable the ad servers to track user preferences and information about what ads they have already seen. Cookies of major web ad servers such as DoubleClick (Google) are also used in serving personalized advertisements in domains other than their own domain. In this context, they are referred to as third-party cookies. Some web advertisers allow users to opt of their personalized advertisements.

Ad servers also track "conversions" resulting from advertisements, which for example can be an online purchase or viewing another web page. Conversion analysis allows the advertisers to evaluate which ads have been effective against which market segments and refine their marketing strategies further. This can be considered a unique aspect of online advertising since coming up with such a measure for television advertising for example that relies on the television medium itself would be impossible.

An initiative named "Do Not Track" is aimed at allowing users to indicate that they do not want to be tracked via an HTTP header field. The initiative has gained some traction in the sense that this mechanism is now supported by most browsers. However, very few websites honor the "Do Not Track" header and it is unclear whether the initiative will gather more support in the future [11].

Web sites that possess a large amount of user information can mine that data for many purposes other than online advertising. A recent article in the MIT Technology Review describes some data mining projects carried out by data scientists at Facebook [12]. It mentions projects aimed at identifying what types of updates from friends encourage newcomers to the network to add their contributions and the songs most popular with people who have recently left a relationship.

Since data of children over the age of 13 are not covered by COPPA, currently web applications have no legal requirement to handle their data differentially in these data mining applications. However, our position is that it is better for the industry and the data mining community to pay special attention to ethical concerns when mining such data. We elaborate on this point in the following section.

The data mining techniques that are used in mining personal data can range from straight-forward association rule mining to specialized techniques aimed at a particular task. For example, Google has several patents on personalized advertising.

3. ETHICAL CONCERNS

This section serves two purposes. Firstly, we summarize our observations on ethical issues on data mining related to children. Secondly, we propose recommendations on the issues discussed. Some of these recommendations are based on the literature and industry practice while others are recommendations of our own. The section is divided into several subsections based on the ethical issues identified.

3.1 Acceptability of the Practice

Due to the seriousness of the charges that have been raised against mining personal data of children, the first ethical question that needs to be settled is whether mining such data is ethical under any circumstance.

COPPA is relevant in answering this question with respect to children under the age of 13. Since it states that personal data of children should be kept only as long as necessary for the purpose they were collected for and that a website cannot condition a child's participation in online activities based on providing more information than necessary, it strongly implies that such data can only be mined if it is relevant to the purpose for which the children joined the application. For example, mining details of a child's participation on an education portal should be acceptable for the purpose of providing a better learning experience. However, using the same data to provide personalized advertisements in the said portal raises ethical and legal concerns for children under the age of 13.

Recognizing that mining personal data of children carried out for purposes such as education can be beneficial to the children shows that mining data related to children cannot be rejected outright as an unethical exercise. However, ethical concerns remain and care should be taken to address them, as shown by COPPA.

The experience of American Psychological Association (APA) in addressing the ethics of psychologists working with the marketing industry provides another perspective on this: because of antitrust concerns, the ethics code of an organization for data mining practitioners, should there be one, cannot probably specifically prohibit the practice of mining children's personal data. It can however provide details on how the general code of ethics applies to this situation.

From a legal point of view, mining data of children over the age of 13 does not need any additional scrutiny when compared with mining data of adults except in special cases such as the release of educational records. However, it is better for the industry to pay special attention to the ethics in such data mining exercises given the social responsibility in protecting children. We return to this perspective in the following subsections.

3.2 Differential Treatment in Educational Settings

The lawsuit against Google for scanning student emails in its Google Apps for Education shows that people may find it objectionable when personal data is mined in an educational setting for a purpose other than education. While the legal arguments used in the case concentrate on possible release of educational records covered by FERPA, from the media coverage on the case it can be seen that the real ethical concern raised here is mining of data collected in an educational setting for commercial purposes.

Therefore, it is reasonable to recognize that mining of data collected from children on an educational setting should be held to a higher standard than mining of data collected in other settings, for example in a social network. A reasonable guideline here is to use such mining only for enhancing the students' learning experience. At least some people are likely to find mining such data for any other purpose objectionable, even when not driven by a direct commercial interest.

The agreement by Google to stop scanning email in Google Apps for Education shows that the industry is already moving in this direction. Such moves will help alleviate the concerns raised on mining children's data in educational settings.

3.3 Transfer of Data

If we are to consider COPPA as a guideline on dealing with data related to children, it can be hypothesized that transfer of data on children from an organization that collected the data to another organization raises ethical concerns. This is because the spirit of the legislation is on using the data collected from children only for the purpose it was collected for. While COPPA applies only to children under the age of 13, its implications are important in a wider context.

Good anonymization techniques reduce some concerns on transferring data. However, research work has shown that it is extremely difficult to ensure that another party will not be able de-anonymize the data by using it in conjunction with other information [13]. This is especially important given that "repurposing data", where data is used for a different purpose than the one it was created, has received a lot of attention in data science and data mining applications. Therefore, an organization that chooses to transfer the data on children collected by it to other organizations is taking a significant risk.

Transfer of data related to children without anonymization is especially ethically challenging in the context of data brokers building up detailed profiles of users through data collected form a large number of data points, as stated in a recent report by the Federal Trade Commission [14]. The current focus is on the use of such profiles for adults, but it can be hypothesized that data collected from children can be used to augment the profiles built for them once they become adults. We return to this topic in Section 3.6.

3.4 Transfer of Mined Knowledge

Transfer of knowledge mined from data on children to other organizations is a different ethical issue from transferring the data itself. As a simple example, rules discovered through association rule mining on children's data in a social network can reveal which characteristics are correlated with preferences for certain products and such rules can be of great value to marketers. If the social network chooses to sell such rules, without the accompanying data (and if marketers agree to purchase such knowledge), this arguably raises less ethical concerns than transferring the data itself.

However, it is important to recognize that even this seemingly benign practice is not guaranteed to be free of ethical concerns. For example, consider the discovery and sale of correlations between fast food usage in teens and their lifestyle choices. Such information can be used in marketing fast food more effectively to teens but given the negative social perceptions towards fast food, this is bound to be viewed negatively.

Therefore, it can be argued that the ethical concerns have to be evaluated on a case-by-case basis even in transferring knowledge mined from children's data. Parallels can be made between these ethical challenges and the ethical challenges faced by psychologists in working with marketers on children: both data mining practitioners and psychologists have to think about the effects of their contributions on children and the wider society.

3.5 Personalized Ads

News articles about personalized advertisements in Google Apps for Education and age-inappropriate ads shown to some children in Facebook [15], which have slipped through its review systems, shows that display of personalized ads to children is a sensitive issue. In the case of inappropriate Facebook ads, pointed out by an article in Wall Street Journal, the company has removed them immediately after the disclosure. On its filters for age-inappropriate ads, the company has stated "no system is perfect. When we find or are made aware of prohibited ads, we remove them immediately".

It is important to note that such concerns have been raised against specific cases of presenting personalized ads for children and not against the wider practice. It appears that there is no widespread calls for prohibit personalized ads for teens. Rather the concerns are on showing inappropriate ads and showing ads in an educational setting.

Therefore, in the case of personalized ads, the important requirement is to have sophisticated filters to weed out inappropriate ads out of the ads selected based on user profiles. In addition, it is necessary to examine whether personalized ads should not be used when children are interacting with an application in a specialized setting.

3.6 Transition from a Child to an Adult

Since web sites and collect data from children over the age of 13, a child would have already created a significant digital footprint by the time he turns 18. Presumably, such data is used together with data produced by him after he becomes an adult by websites that he used as a child and continues to use as an adult.

Legally, this should not be a problem. However, it is not difficult to come with scenarios where this practice would result in undesirable outcomes, especially in the context of data brokers gathering a lot of information about people from different data points. For example, the choice of books read as a child might reflect negatively on a person, if market segments are identified based on this factor. While classifying people based on personal information may give rise to ethical concerns on its own, grounding such classifications based on actions of a person as a child would raise even more concerns.

Therefore, it can be argued that care should be taken in building profiles for users combining data from their childhood and adulthood. Being aware that this can give rise to ethical issues will be helpful in preventing them.

4. CONCLUSION

Our position on mining data related to children, presented in the previous section, can be summarized as follows: it should be considered acceptable, provided that special care is taken in the exercise. The data produced in an educational setting should be held to a differential, higher standard. Transfer of data from the organization that collects them to another organization is not acceptable, even with anonymization. Transfer of knowledge mined for data is acceptable, but care should be taken in this exercise. Personalized ads are also acceptable but with smart filters to leave out inappropriate ads and with the caveat that they may not be acceptable in specialized settings. Special care should be taken in combining data produced by the same person as a child and as an adult.

In a wider context, we believe that paying close attention to ethical issues associated with data mining, especially when dealing with data related to people, is essential to the continued development of the data mining industry and the research community. Ethics associated with data mining are receiving increasing attention from the general public and government organizations (such as the Federal Trade Commission) and as such the data mining community should respond to the ethical concerns raised. In addition, it should anticipate ethical issues that may arise in the future and start addressing them. Our work is aimed at making a contribution towards this broader objective.

5. REFERENCES

[1] Simon, S. 2014. Data mining your children. *Politico* http://www.politico.com/story/2014/05/data-mining-your-children-106676.html

[2] Hern, A. 2014. Google faces lawsuit over email scanning and student data. *The Guardian*. http://www.theguardian.com/technology/2014/mar/19/google-lawsuit-email-scanning-student-data-apps-education

[3] Ontario Secondary School Teachers' Federation. 1995. *Commercialization in Ontario schools: a research report*.

[4] Wikipedia. *Children's Online Privacy Protection Act*. Accessed May 29, 2014. http://en.wikipedia.org/wiki/Children's_Online_Privacy_Protection_Act

[5] Harrison, P., Chalmers, K., d'Souza, S., Coveney, J., Ward, P., Mehta, K., and Handsley, E. 2010. *Targeting children with integrated marketing communications*. Flinders University.

[6] McNeal, J.U. 1992. *Kids as customers: a handbook of marketing to children*. Lexington Books.

[7] Beder, S. 1998. A community view: caring for children in the media age. In *Papers from a national conference*, New College Institute for Values research, pp 101-111.

[8] Clay, R.A. 2000. Advertising to children: Is it ethical. *Monitor on Psychology*, 31, 8, p 52.

[9] Wilcox, B.L, Kunkel, D., Cantor, J., Dowrick, P., Linn, S., and Palmer, E. 2004. *Report of the APA task force on advertising and children*. American Psychological Association.

[10] Kramer, J.B. 2006. Ethical analysis and recommended action in response to the dangers associated with youth consumerism. *Ethics and Behavior*, 16, 4, pp 291-303.

[11] Miners, Z. 2014 How bickering and greed neutered the 'Do Not Track' privacy initiative. http://www.pcworld.com/article/2158220/do-not-track-oh-what-the-heck-go-ahead.html

[12] Simonite, T. 2012. What Facebook knows. *MIT Technology Review*, 115, 4, pp 42-48.

[13] Sharad, K. and Danezis, G. 2013. De-anonymizing D4D datasets. In: *Proceeding of the Workshop on Hot Topics in Privacy Enhancing Technologies*.

[14] Federal Trade Commission. 2014. Data brokers: a call for transparency and accountability.

[15] Elder, J. Nude webcams and diet drugs: the Facebook ads teens aren't supposed to see. *The Wall Street Journal*. http://online.wsj.com/news/articles/SB10001424052702304703804579381552745011772