Proposal for a “Down-the-Chain" Notification Requirement in Online Behavioral Advertising Research and Development

David Thaw
Department of Computer Science
University of Maryland College Park
dbthaw@cs.umd.edu

Neha Gupta
Department of Computer Science
University of Maryland College Park
neha@cs.umd.edu

Ashok Agrawala
Department of Computer Science
University of Maryland College Park
agrawala@cs.umd.edu

1. INTRODUCTION

Contextual online advertising, also known as behaviorally-targeted or “Online Behavioral Advertising,” is one of the fastest growing markets on the Internet [4, 1, 6]. These terms are used to describe a variety of Internet advertising services, many of which collect information about individuals’ identity, personal characteristics, preferences, and online behaviors. Marketers consider such information quite valuable and use it to deliver advertising content on an individual basis. By customizing the delivery of advertising content, marketers argue that consumers receive information more relevant to their individual needs and wants [7].

Much of the information collected, however, may be of a sensitive nature and/or may conflict with consumers’ privacy expectations [13]. There also is substantial confusion among consumers as to what information is collected and how that information is used. As noted by many scientists [14] and policymakers [2], privacy policies (in their current form) are ineffective at informing consumers of what information collection occurs, what options are available regarding such collection, and how consumers would go about exercising those options.

In this paper, we propose a system called “down-the-chain” notification, under which producers at each step of the research, design, implementation, and maintenance stages bear the responsibility to document the information input *“needs”* of their algorithms and other technical elements to ensure accurate information is available as to the actual needs of the technology. We feel that such a requirement will help improve consumer options and help ensure those choices are enforceable.

2. CONTRIBUTION

As computer scientists and attorneys working in the behavioral advertising space, we feel that informed consumer choice is essential to the continued viability and vibrancy of the Online Behavioral Advertising (OBA) industry. We believe this workshop is an important opportunity for technical policy and business stakeholders to interact. We seek feedback on our proposal and we hope to use this workshop as an opportunity to engage the input of these stakeholders to improve our proposal.

For consumers to be informed effectively, the drafters of consumer notices must have access to complete and accurate descriptions of what information is being collected about individuals and how that information is being used. Likewise, in making decisions regarding the design and implementation of systems, business and technical staff must have access to complete and accurate information describing the data needs of various algorithms and other technical components upon which the systems they implement are based.

This flow of information from the whiteboards where algorithms are first conceived to the end-user/consumer via privacy policies and other notice-and-choice mechanisms is essential to ensuring that:

- The “administrators” of Online Behavioral Advertising systems (e.g., ad networks wishing to collect data and perform analytics) have the maximum number of consumer (privacy) choices available to offer consumer users of the content (e.g., website visitors); and
- The published consumer expectations (e.g., via privacy policies) are as accurate as possible and not erroneous as a result of disconnects between the attorneys and privacy professionals drafting notices and the technical developers actually implementing the design decisions in software.

As experienced computer scientists, we are aware of the challenges inherent in pressing for any standard requiring developers to document their code. However, given the heightened privacy concerns inherent in this space, and the notable gap in technical understanding between those individuals drafting consumer-facing materials and those individuals designing/maintaining the systems, we believe that in consideration with the potential sensitivity of the information at issue, a higher standard is in order.

Our paper discusses these issues in greater depth, using as an example research currently being conducted at Maryland. We propose a framework for this ”down-the-chain” notification and raise several issues for discussion that we feel are still outstanding in our position/proposal.

3. DOWN-THE-CHAIN NOTIFICATION

We propose that those who design and implement the technologies enabling OBA have a responsibility to document the “information requirements” of their technologies. Adopting a well-known principle in software engineering called precondition/postcondition documenting, designers and implementers of OBA systems would be required to specify what types of information must be collected and what types of information must be persistently stored for each function of their system to operate.

Since there are different types of (often sensitive) information involved in the OBA system, we suggest a “down-the-chain” notification system where entities involved in any of the five roles as mentioned in Section 3.1, no matter whether they function independently or collaboratively, communicate the information needs of their work to the next role in the chain. So, for example, when computer scientists propose new algorithms they must also state the requirements those algorithms have with regard to what information must be collected and what information must be persistently stored for the algorithm to function. This helps designers building the actual systems make better-informed decisions about what information to keep and what can be discarded (or not collected at all). Moving down the chain, the work of the business professionals and privacy professions becomes easier due to this efficient system of documenting the data requirements at each step.

The goal of this requirement is two-fold. First, to enable choice by allowing decision-makers (business professionals and consumers) to choose not to collect/retain or not to supply any more information than necessary for operation of the system. Second, to help enforce choices by providing visibility into exactly what types of information are used. We explore how this requirement might function in the context of a system currently under development at the MIND Lab at the University of Maryland [3].

3.1 Roles

There are multiple entities or players involved in the design, implementation and usage of OBA system. We discuss this notification requirement in the context of five generalized roles: 1) computer (and other technical) scientists; 2) software engineers and system administrators; 3) business professionals; 4) attorneys/privacy professionals; and 5) end-users/consumers. We define these roles as follows:

- **Computer (and other Technical) Scientists:** Those who design the fundamental algorithms and other technical elements upon which OBA systems are based.
- **Software Engineers and System Administrators:** Those who design and maintain the information systems that support Online Behavioral Advertising.
- **Business Professionals:** Those who make business decisions according to the requirements of OBA systems.
- **Attorneys/privacy professionals:** Those who draft consumer communications and vet the requirements of OBA systems and software for legal and regulatory compliance.
- **End-Users/Consumers:** The *users* of OBA systems and the websites supported by OBA advertising revenue. These individuals should be able to make informed choices as to their participation in and use of OBA systems and the websites and other technologies OBA revenue supports.

In proposing a down-the-chain notification requirement, our goal is to ensure two conditions:

- The maximum number of choices are made available to each decision-maker in the chain; and
- The choices made are both given effect and are verifiable (auditable).

3.2 Types of Information

In certain cases, for example the protection of proprietary trade secrets regarding system design, it may be desirable to allow developers to report what information must be collected/stored in categories rather than precise (individual) elements. This approach also allows for flexibility in upgrades/modifications if a new data field is added (that does not fundamentally change the privacy landscape). Additionally, this approach may serve to simplify the presentation of the information collection/retention specifications.

To give this approach effect, we propose segregating data elements into the following categories:

3.2.1 Static Demographic Information

Static Demographic information is the demographic information such as gender, age, marital status, family size, user-defined interests, race/ethnic origin and the genetic make-up, religion, etc. This static demographic information does not change in the short term or with user “browsing behavior”.

3.2.2 Personal Information

Information explicitly capable of uniquely or near-uniquely identifying an individual:[1]

- Uniquely identifiable information: includes full name[2],

[1] The categories that follow list example data elements and are not intended to convey comprehensive lists.

[2] Assume full name collisions are resolvable.

SSN, mobile phone numbers[3], financial account information[4].

- Near-uniquely identifiable information: includes landline phone numbers, email addresses, mailing addresses.

3.2.3 Behavioral Information

Any type of past browsing actions including user engagements such as mouse overs, non-navigation clicks, etc. along with the search queries issued by the user is called behavioral information. Additionally, if knowable, the amount of time spent in any of these activities is called behavioral data.

3.2.4 Modeled Information

Modeled Information consists of generalizations about interests, behaviors, etc. based on past behavioral events but without including any specific behavioral events. Modeled information usually includes predictions and rules made by humans or inferencing algorithms over the behavioral and static demographic data.

3.3 Inferencing Algorithms

Inferencing algorithms are used to make sense of these vast amounts of data or information collected in OBA systems. Different algorithmic techniques use different feature sets of the data and also have different requirements on the storage time of the data. We can broadly classify the learning algorithms into two categories: offline learning and online learning.

- **Offline Learning** : The traditional method for doing inferencing where models and rules are developed and updated from (static) historical data and then are applied to future data.
- **Online Learning**: A dynamic method of updating the learnt models in real time as new data is encountered.[5]

4. ARCHITECTURE OF AN EXAMPLE ONLINE ADVERTISING SYSTEM

To illustrate our “down-the-chain” system, we present here an architecture which will enable developers, designers and system architects to maintain a coherent view of the OBA system as shown in Fig. 1. This example architecture demonstrates how data can be segregated to restrict access to individuals’ personal information while still enabling certain personalization features of OBA.

It is generally preferable from a software engineering perspective to implement privacy features while the system is being designed rather than attempting to “layer” those features on to an existing system[5, 12]. To make informed design decisions, software engineers must be able to identify what information is required, when the information is required and what elements of the system require the information.

To ensure that the above principles are followed, we divide the system into 3 parts:

- **Globally Unique Identifier (GUID) Table and Personal Information(PI)** : GUID table contains a mapping of username to globally unique identifier which can, for example, be stored in a browser cookie. All other databases are keyed to the GUID. PI is the database containing personal information such as username, full name, passwords, etc. The access to personal information is restricted and “firewalled” off from the rest of the system. PI is a write-only database to prevent behavioral and other data from being linked to individuals’ explicit identity. We consider the PI database to be write-only in the context of our example system for the purposes of this paper. We recognize that in practice this write-only state will be enforced by a matter of policy, and that reads of this database are necessary for other functions (e.g., regulatory compliance with COPPA[6]).
- **Demographic and Behavioral Information**: Demographic and Behavioral information can be stored as part of the reporting and logging system and time to time inferencing can be performed over the data depending on how the system is set up. After each impression[7], demographic and behavioral information can be updated for the GUID associated with the impression. If online learning is performed then requisite information from these databases is passed on to the modeled information knowledge base.
- **Inferencing Algorithm and Modeled Information**: As discussed in Section. 3.3, the inferencing algorithms can be of various types and have different information needs. Hence, modeled information can vary depending on the algorithm needs. Estimating the needs upfront always helps both the developers and the policy makers.

4.1 Example

As an example of the implementation of the “down-the-chain” notification policy, we talk about a system proposed by our research group at the University of Maryland. Our role in the OBA system is that of a *computer scientist* developing the fundamental algorithms. We specify our information needs at each phase of the algorithm which has made it easier for the actual implementors and designers to design an efficient system and to make available as many privacy options as possible for their system design.[8]

[3] Mobile phones may have many-to-one numbers-to-individuals mapping, however this is much less frequent than with email addresses and landline phone and thus we place mobile phones in the uniquely-identifiable category.

[4] Treat joint accounts as single person for the purposes of discussion.

[5] There are also batch-learning algorithms which lie in between of online and offline learning.

[6] Children’s Online Privacy Protection Act, 15 U.S.C. §§ 6501-06.

[7] A single view of a webpage is called an impression.

[8] This approach does not propose that roles earlier in the chain *require* various privacy features, only that they enable as many as possible and accurately document information needs so as to make the next role in the chain aware of the available options.

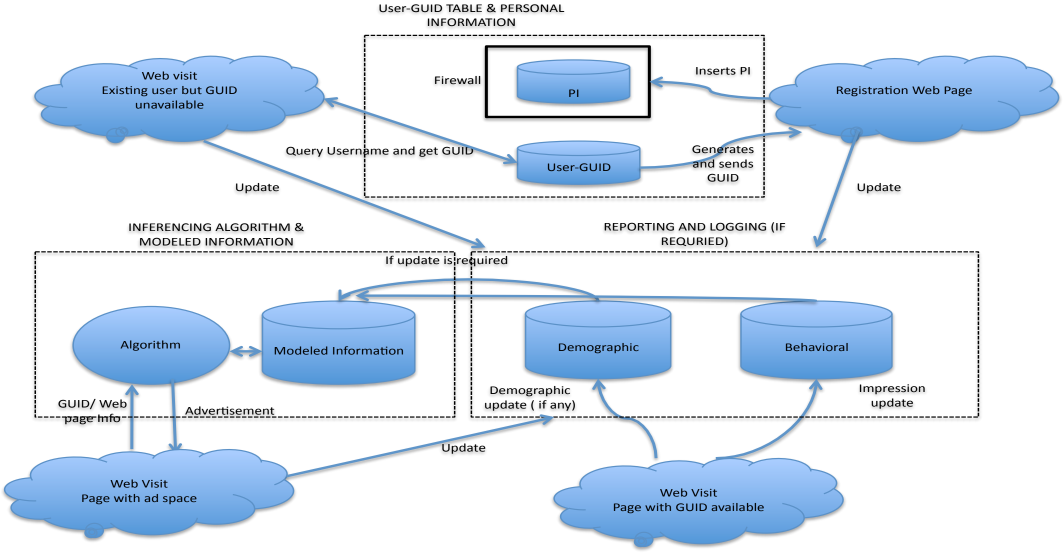

Figure 1: Sample Architecture for Online Advertising Server. PI refers to Personal Information and GUID is the Globally Unique Identifier.

Ad_1	$\#clicks$	$\#impressions$
Ad_2	$\#clicks$	$\#impressions$
...	...	...
Ad_n	$\#clicks$	$\#impressions$

Table 1: Modeled Information Phase 1

The online advertising system suggested by our group uses a form of "*Contextual Bandit Algorithm*" [10]. Contextual Bandit Algorithms perform online learning and are an advanced form of the multi-armed bandit problem [8]. Our algorithm is being developed incrementally and we have divided its development into 3 phases.

4.1.1 Phase 1: Multi Armed Bandit Problem

Our system uses a bayesian inferencing based multi-armed bandit formulation to model the problem of online optimization in advertising. The multi-armed bandit problem can be formulated as follows: there is a bandit containing a set of arms $(A_1, A_2, A_3, ..., A_n)$. Each arm has a success probability θ_i associated with it which is unknown. A strategy needs to be decided to play the arms such that the total rewards obtained are maximized and the learning cost is minimized.

In the advertising domain, we can model each advertisement as an arm of a bandit and each impression of an advertisement as a play of an arm. The goal is to maximize the total number of clicks obtained by presenting more attractive advertising. Different multi-armed bandit strategies have been discussed in [11, 9]. These strategies are agnostic to the user and its information, hence the behavioral and static demographic data is not required. The only piece of information required for inferencing are the advertisement characteristics:

$Segment_1$		
Ad_1	$\#clicks$	$\#impressions$
Ad_2	$\#clicks$	$\#impressions$
...	...	...
Ad_n	$\#clicks$	$\#impressions$

.....

$Segment_k$		
Ad_1	$\#clicks$	$\#impressions$
Ad_2	$\#clicks$	$\#impressions$
...	...	...
Ad_n	$\#clicks$	$\#impressions$

Table 2: Modeled Information Phase 2

In the current form, there is no need for the system to store any information indexed or indexable to the user and only needs to store per advertisement information as shown in Table 1. The system will work fine if the necessary counters are incremented appropriately and the rest of the data is discarded (or not collected).

4.1.2 Phase 2: Segmented Multi Armed Bandit Problem

It is a common understanding that different users behave differently and adding the user behavior can provide a boost in the advertising systems. Hence users can be segmented into groups and a separate model can be built for each group of the users. In our model, we use historical data consisting of demographic as well as behavioral data for segmenting users and then build a separate multi-armed bandit model for each segment as shown in Table 2.

When a new impression arrives with the GUID, a simple look-up is done to find out which segment the user belongs to and then the appropriate model is used. Otherwise, when a new user comes his information is updated in the behavioral and demographic data and then his segment is decided using a pre-calculated formula for the segmentation.

Hence, in this phase the system needs to store only the information mentioned in Table 2 and the formula (mapping) for user segmentation. All the other behavioral and demographic information can be discarded after the segmentation has been done.

4.1.3 Phase 3: Contextual Bandit Problem

The preferences of the users change with time and dynamic systems can more accurately model individuals' preferences. So in its third phase, our system will also update the user level segmentation in an online fashion with each impression.

Much more information will need to be maintained and updated in this model since user context will be taken into account in an online manner in deciding which advertisement to display. [9]

Since *computer scientists* are one of the first in this chain of notification, we play an important role in fulfillment of the requirement of "down-the-chain" methodology.

5. OPEN QUESTIONS FOR DISCUSSION

This proposal represents preliminary discussions in the MIND Lab regarding the responsibilities we bear as computer scientists when developing new technologies. We have identified some questions that remain open issues in our proposal:

- Should developers also have to make public the data requirements (not just make that information available to purchasers/users of their systems)? Should these specifications be auditable?
- If there are different entities which supply different information pieces, then how should these information items be stored and managed? (e.g., if each of Yahoo and Google supply inputs to a behavioral marketer, should the marketer be required to publish the requirements of both systems?)
- Should advertisers (as different from operators of Ad Networks) have a role in the notification process?

6. CONCLUSION

We advocate that those who design and implement Online Behavioral Advertising systems should be required to document the information requirements of the algorithms and other technologies they build and maintain so that it is possible both to build systems that do not store more information than necessary and to enable others to audit whether more information is being stored than necessary. The main goal of this requirement is to enable as much consumer choice as possible.

Our *"down-the-chain"* notification policy will bridge the gap between the understanding of the functionality & requirements of the system amongst different role players and will make it easier for each player to work on his part. It will also provide a more coherent and useful view to the *end-consumer.*

7. REFERENCES

[1] http://advertising.yahoo.com/.

[2] http://mediadecoder.blogs.nytimes.com/2009/08/05/an-interview-with-david-vladeck-of-the-ftc/.

[3] http://mindlab.umd.edu/.

[4] http://www.google.com/ads/.

[5] http://www.privacybydesign.ca/.

[6] http://www.teracent.com/.

[7] Consumers driving the digital uptake: The economic value of online advertising-based services for consumers. Interactive Advertising Bureau (IAB) Report, September 2010. http://www.iab.net/insights_research/947883/.

[8] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite time analysis of multi-armed bandit problem. *Machine Learning*, 27(2-3):235–256, 2002.

[9] D. Chakrabarti, R. Kumar, F. Radlinski, and E. Upfal. Mortal multi-armed bandits. In *NIPS*, 2008.

[10] L. Li, W. Chu, J. Langford, and R. E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *WWW*, 2010.

[11] S. Pandey and C. Olston. Handling advertisements of unknown quality in search advertising. In *NIPS*, 2006.

[12] S. S. Shapiro. Privacy by design: Moving from art to practice. *Communications of the ACM*, 53(6):27–29, 2010.

[13] J. Turow, J. King, C. J. Hoofnagle, A. Beakley, and M. Henessy. Americans reject tailored advertising and three activities that enable it, September 2009. SSRN: http://ssrn.com/abstract=1478214.

[14] F. Williams. Internet privacy policies: A composite index for measuring compliance to the fair information principles, 2006. http://www.ftc.gov/os/comments/behavioraladvertising.

[9] We are still working on the exact details of this system which will determine the sufficient information elements.