Decentralized recommender systems for mobile advertisement

Andrés Moreno
School of Engineering,
University of los Andes
Cra 1 N°18A- 12
Bogotá, Colombia
dar-more@uniandes.edu.co

Harold Castro
School of Engineering,
University of los Andes
Cra 1 N°18A- 12
Bogotá, Colombia
hcastro@uniandes.edu.co

Michel Riveill
Laboratoire I3S (Université de Nice Sophia Antipolis)
930 route des Colles BP 145
Sophia Antipolis, France
riveill@unice.fr

ABSTRACT

Advertisement in mobile devices is a key activity for the monetization of mobile applications. Behavioral targeting is a preference elicitation technique currently used in online and mobile scenarios for presenting users ads that are highly related to their interests. Although successful, behavioral targeting gathers personal information about users into a centralized entity which raises several privacy concerns. To solve these problems recent approaches for preserving privacy rely on a client-side profile elicitation technique and storage. In this position paper we argue that these proposals might be limiting the accuracy of local based profiles and we argue for revisiting decentralized recommender systems as a mean to develop effective user profiling strategies for mobile advertisement purposes while keeping the privacy of users in the system.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information filtering; K.4 [**Computers and Society**]: Public Policy Issues—*Privacy*

Keywords

mobile advertisement, privacy, p2p

1. INTRODUCTION

The mobile application market is expected to have a steady growth through 2014 as the number of downloaded applications will increase from 10.9 billion worldwide in 2010 to 76.9 billion in 2014 [11]. The most common way for developers to obtain revenue from the development of these applications are:(1) making paid applications for application stores (*app-stores*) and (2) embedding third party advertisement messages (ads) into slots available inside the applications provided by advertisement networks such as Google's AdMob and Apple's iAd. According to recent figures [16] the later approach has been gaining importance since customers are more likely to download free equivalents of paid applications than paying for them, particularly in Asian markets. The adequate selection of ads to show to adequate audiences in mobile applications represents a challenge for advertisement networks as developers are progressively turning to advertisement as their most important mean for monetization.

Advertisement networks use Behavioral Targeting (BT) techniques [5] to choose which ads should be displayed for the user. BT is a profiling technique initially used in online advertisement that learns user's interests by mining information about the past behavior of users. BT gathers information such as their browsing navigation history and queries made to a search engine in order to increase the click-through rate of served ads.

This technique has been proven successful [5] [19] in online advertisement, however the collection and storage of detailed information of user behavior has raised some privacy concerns. Advertisement networks are being trusted with a great amount of personal information and users have limited action to control how their data is used, and even sometimes they are not aware that these networks are gathering sensitive information about themselves [6].

Privacy concerns increase when talking about BT in mobile environments. We believe that the popularization of mobile devices and the capabilities of connectivity to the Internet of these devices allows advertisement networks to perform a narrower personalization and preference elicitation in a mobile scenario when compared to the online scenario since: 1) Mobile devices are rarely shared between different users, contrary to online advertisement where different users can be behind the same IP or where different users can share the same computer. Since the probability of multiple users using the same device to interact with Internet services is low, targeting techniques are confident to assign all the detected interactions of the device to a single user profile; and 2) Capabilities of modern mobile devices such as GPS or network-based location allow applications to know the exact location of the user, permitting advertisement networks to deliver ads highly related not only to the user's interest but also highly related to the user's location. As recognized

by [12], the narrower understanding of the user situation by third party organizations in the mobile scenario ultimately aggravates the privacy concerns that currently exist in online advertisement.

In the rest of this position paper we will review the privacy concerns that exist in these kind of systems, the related initiatives to protect privacy of users, and our proposal for a new research direction to address the problems mentioned beforehand.

2. PRIVACY CONCERNS OF ADVERTISEMENT SYSTEMS

Advertisement networks systems gather information about users and store it in a centralized entity. Then they apply data mining techniques to learn the users' interests with the purpose of showing ads that are relevant. On the other hand, by downloading an application from an app provider users entrust personal information to the advertisement network expecting in return access to mobile applications and eventually to be led to good commercial opportunities with the ads presented to them. The tension between the need of advertisement networks to target users and user's desire to keep their information private while keeping the benefits of targeting results in a *personalization-privacy paradox* that is inherent to this type of systems.

The trust that users put on advertisement networks is an agreement that they will use personal information only for purposes mentioned beforehand, however when advertisement networks consolidate user information into a centralized entity they increase the risk that this information is used for purposes different from mobile advertisement. This *exposure* risk can be configured in five ways [7] :

- *Deception by the recipient:* The system can lie about its privacy policies and trick users to reveal personal information, using it later for a different purpose from profiling for advertisement display. For example selling the information or sharing it with other organizations.
- *Mission creep:* Initially the policy of usage of personal information is defined clearly by the system, but later the systems expands its goals in a previously unforeseen manner, changing the use of personal information for other purposes related to the new goals of the organization.
- *Accidental disclosure:* Information about users can be made available accidentally, for example leaving private information on a server that can be accessed by a search engine over the Internet.
- *Disclosure by malicious intent:* Storage servers' security can be breached and users' personal information can be stolen.
- *Forced disclosure:* Systems must disclose the information for legal reasons.

When personal information is exposed, users are subject of various potential harms [12] for example: 1) Users can be targets of unwanted commercial solicitations (spam); 2) Users can be victims of identity theft and fraud ; and 3) Exposure of personal information increases the user's risk to be subject of unfair commercial practices. This harms are more critical in mobile advertisement where the certainty of the information gathered about the user is higher.

3. RELATED WORK

In order to avoid *exposure* risks, recent approaches for online and mobile advertisement hide user information from the advertisement networks by delegating the profile elicitation techniques to computational agents that run on the user's device, preventing the advertisement network from gathering information about the user and reducing the exposure risks presented in the previous section. This represents a challenge because advertisement networks need user information for other processes beside user profiling that are crucial to their business model. For example advertisement networks need to know which ads are displayed to users and which of those ads are clicked on so they can know how much to charge advertisers and how much they have to pay to developers that allow ads in their applications. Other reason they need this information is to defend themselves when a malicious client clicks on ads in order to increase artificially the billing of advertisers or the revenue perceived by application developers.

The *Privad* system [8] is a online advertisement platform that learns a local user profile . In this system the client-side agent requests ads to an intermediate broker by submitting an incomplete user profile with some general categories of interests among with some demographic information about the user. The broker has a local database of ads from advertisement network's and does a coarse-grained selection of the most relevant ads for the user based on the profile submitted by the client-side agent. The client-side agent receives a list of possible ads and does a fine-grained filtering on the list using the complete local profile, choosing the most relevant message to be displayed for the user. Unused ads are cached for future use. The client agent reports ad events (which ads have been displayed and clicked on) to the broker through an anonymizing proxy called dealer for billing and ad-click fraud detection purposes. The *Adnostic* system [17] also keeps a local user profile and works in a very similar way to the Privad system, however in the Adnostic system no anonymization of client-ad interaction is applied in order to facilitate the current work of advertisement networks.

The *MobiAd* [10] system is another platform that learns a user profile based on the different sources of information located inside the user's mobile device such as their local browser history, information from social networks and email messages. This system also takes advantage of the location capabilities of the mobile device and integrates into the user profile the preferred locations and mobility patterns of the user. The focus of MobiAd is to display ads that are highly related to their location. In order for the client to obtain ads without disclosing the user's location, it uses an anonymous broadcast platform to obtain all the ads that are relevant for a specific location and then screens out the ones that are not relevant based on the attributes of the local profile. To report back ad events to the advertisement network, the client uses a delay tolerant network that anonymises the user-ad interaction.

Generally speaking, the main contribution of the approaches presented in this section is that the proposed architectures fit to current advertisement network business models while keeping the user's privacy by keeping from a central authority user detailed user profile data. However, to the best of our knowledge, little or no information is provided about the ability of the proposals to accurately portray the user's interests by only using the local information present in the device. Profiling at client-side relying only on the local history of the user without taking into account the trends of choices of other users of the system is analogous to the strategy employed by *content-based filtering systems* (CB). CB filtering systems are information filtering systems that use only information about past choices of the user in order to build a user profile. This strategy has several limitations as noted by researchers like [1], in particular *Overspecialization*: A CB system will only classify as relevant ads that are similar to the ones that the user has liked in the past. This means that the system can't predict correctly the relevance of an ad that is very different from the ones that the user has seen in the past or belongs to a topic in which the user has not manifested an opinion yet.

A complementary approach to CB systems are *collaborative filtering systems* (CF). CF systems predict the relevance of an item (ads) based only on the opinion that other users have had on items , therefore being completely agnostic to the content of the data item. CF systems can be classified into two classes: 1) The neighborhood CF system that calculates the relevance for a data item by taking into account the opinion on that item of the N most similar users(*user-based CF*) or by taking into account the opinion of the user to the most similar data items to that item (*item-based CF*); and 2) model based CF that learns a latent model from the users' interaction with the system. These systems are not vulnerable to overspecialization since they can detect an accurate relevance for data items that are different from the ones the user has seen. Model based techniques have shown better results than neighborhood approaches in several scenarios, however the calculation of the predictive model requires computational resources provided by servers or by cloud computing architectures, thus they not suitable for profiling in agents with limited computational capabilities like the ones present in the reviewed architectures. Other shortcoming of CF systems is their adaptation to scenarios where the underlying data item set (in this case the ads) is very dynamic and items appear or disappear frequently, for example in [9] it was estimated that between 30 and 40% of available ads in an ad network change from hour to hour. Lastly we want to note that applying this strategy to BT systems with client-side profiles breaks the rule of not sharing the profile with other entities, thus increasing the risk of exposure of personal information of users.

Other kind of preference elicitation systems are hybrid systems. These systems combine different content based and collaborative paradigms in order to overcome the limitations of pure content based and collaborative techniques. Since collaborative filtering is the most popular technique most of the work in hybrid systems has been oriented towards adding features of data items into a collaborative approach. One of the first approaches for hybridization was the *collaboration via content* [14] paradigm. This paradigm uses a neighborhood based collaborative strategy that finds an aggregate user profile of features that represents the knowledge of the user's neighborhood. This aggregate vector is used to complete the opinion on preferences for which the user has not manifested an opinion yet. Other approaches [3] apply a collaborative strategy when there is enough information in the neighborhood of the user to calculate the relevance, if there is not enough information a content-based relevance is calculated. Since hybrid systems use CF at their base, they have also the same shortcomings when trying to apply them on BT with client-side profile elicitation.

4. RESEARCH PROPOSAL: DECENTRALIZED HYBRID SYSTEMS

As mentioned in the preceding section, client-side profiling has limitations that lead to overspecialized systems. On the other hand application of collaborative and hybrid approaches on client-side profiling while keeping the user privacy is not straightforward because: 1) model based CF demands high computational resources not available on mobile devices, 2) neighborhood based CF is not suitable for scenarios where there is a high item churn, and 3) CF is in conflict with the motivation of client-side profiling that recent approaches use.

Privacy concerns have been also recognized in CF systems, there are two trends in CF filtering to preserve the user's privacy: 1) Data obfuscation techniques and 2) decentralization of the user profile database. Data obfuscation of the user profile has been implemented in [15] by randomly modifying values in the users profile on a CF system, this work showed that the impact of the randomization was negligible on the accuracy of the filtering system; however data obfuscation is not sufficient to avoid the risks of information exposure such as deception by the recipient because the database of user-item profiles is still centralized.

Decentralized CF was introduced by [18] by proposing a user-based CF in a p2p environment for portability concerns. Analogous to BT client-side profiles, in a p2p environment each user has a client agent that is in charge of keeping its own profile. In order to calculate the relevance of a new data item, the p2p agents exchange information about the user profiles and keep a local version of other user profiles; once sufficient information about other user profiles is gathered a local version of the filtering algorithm is executed. Decentralized p2p approaches have been implemented using different kinds of CF algorithms: In [4] a user-based neighborhood approach was used; in order to share information with other peers while keeping the privacy of the user, a user profile obfuscation strategy was used. In [13] a decentralized item-based neighborhood was implemented in a p2p architecture among others, the system protected the user privacy by sharing with peers only the similarity index between items according to their personal history. Although these works have embraced the privacy concerns of users, they are based on neighborhood approaches that are susceptible in scenarios where there is a high item churn, specially on [13] where an item-based neighborhood is used.

One of the main concerns of researchers of decentralized CF was the impact on accuracy when operating on a incomplete database of user and item profiles, however [2]

showed that having a small number of peers chosen randomly (roughly 20% of the complete user-item profile information) its good enough to produce comparable results as if the whole database was used. Another concern is how to assure that peers in the system wont use user profile information for other purposes. One promising technique to reduce this risk is to share information only with trusted peers [20], this proposal also has shown that hybrid approaches based on collaboration via content, which are not prone to the high item churn problem, can be used in decentralized environments.

5. CONCLUSION AND FUTURE WORK

In this paper we have reviewed recent proposals for privacy protection of user information in mobile advertisement. Although the proposed architectures fit to the current model of mobile advertisement, client-side profiling is limited and prone to overspecialization. To solve this problem we believe that collaborative techniques should be applied to client-side profiling. We revisited the work that has been done in the recommender system community where the privacy problem has been also recognized, a solution for this problem was the introduction of decentralized recommender systems. In this systems a p2p architecture is deployed to avoid the presence of a central entity gathering all the information of the system. Future challenges in this line of work are: considering the overhead of introducing a p2p overlay network into the system since more computing resources for managing the network can be too demanding for mobile devices with battery life and network limitations, verifying the viability of hybrid systems since the dimensionality of features that describe the data items can be a limitation for the application of this system in a mobile device. Other alternatives that delegate the processing to specialized frameworks such as *personal clouds* can be considered for the task as well, however new challenges such as latency and the risks of delegating user data to a central entity to the user's privacy must be adressed.

6. REFERENCES

[1] G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Trans. on Knowl. and Data Eng.*, 17(6):734–749, April 2005.

[2] A. Bakker, E. Ogston, and M. van Steen. Collaborative filtering using random neighbours in peer-to-peer networks. In *CNIKM '09: Proceeding of the 1st ACM international workshop on Complex networks meet information & knowledge management*, pages 67–75, New York, NY, USA, 2009. ACM.

[3] M. Balabanović and Y. Shoham. Fab: content-based, collaborative recommendation. *Commun. ACM*, 40:66–72, March 1997.

[4] S. Berkovsky, Y. Eytani, T. Kuflik, and F. Ricci. Enhancing privacy and preserving accuracy of a distributed collaborative filtering. In *Proceedings of the 2007 ACM conference on Recommender systems*, RecSys '07, pages 9–16, New York, NY, USA, 2007. ACM.

[5] Y. Chen, D. Pavlov, and J. F. Canny. Large-scale behavioral targeting. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, pages 209–218, New York, NY, USA, 2009. ACM.

[6] D. S. Evans. The online advertising industry: Economics, evolution, and privacy. *Social Science Research Network Working Paper Series*, Apr. 2009.

[7] L. N. Foner. *Political artifacts and personal privacy: the yenta multiagent distributed matchmaking system.* PhD thesis, 1999. AAI0801075.

[8] S. Guha, B. Cheng, and P. Francis. Privad: Practical privacy in online advertising. In *Proceedings of the 8th USENIX Symposium on Networked Systems Design and Implementation (NSDI)*, Mar. 2011.

[9] S. Guha, A. Reznichenko, K. Tang, H. Haddadi, and P. Francis. Serving Ads from localhost for Performance, Privacy, and Profit. In *Proceedings of the 8th Workshop on Hot Topics in Networks (HotNets)*, New York, NY, Oct 2009.

[10] H. Haddadi, P. Hui, and I. Brown. MobiAd: private and scalable mobile advertising. In *Proceedings of the fifth ACM international workshop on Mobility in the evolving internet architecture*, MobiArch '10, pages 33–38, New York, NY, USA, Sept. 2010. ACM.

[11] IDC. Idc forecasts worldwide mobile applications revenues to experience more than 60% compound annual growth through 2014. [Online]http://www.idc.com/about/viewpressrelease.jsp?containerId=prUS22617910, dec 2010.

[12] N. J. King and P. W. Jessen. Profiling the mobile customer - privacy concerns when behavioural advertisers target mobile phones - part i. *Computer Law & Security Review*, 26(5):455–478, Sept. 2010.

[13] B. N. Miller, J. A. Konstan, and J. Riedl. PocketLens: Toward a personal recommender system. *ACM Transactions on Information Systems*, 22(3):437–476, July 2004.

[14] M. Pazzani and D. Billsus. Learning and revising user profiles: The identification of interesting web sites. *Machine Learning*, 27(3):313–331, June 1997.

[15] H. Polat and W. Du. SVD-based collaborative filtering with privacy. In *Proceedings of the 2005 ACM symposium on Applied computing*, SAC '05, pages 791–795, New York, NY, USA, 2005. ACM.

[16] G. J. Spriensma. App distribution becomes a global game the shift of power and impact for developers. Technical report, Distimo, 2011.

[17] V. Toubiana, H. Nissenbaum, A. Narayanan, S. Barocas, and D. Boneh. Adnostic: Privacy preserving targeted advertising.

[18] A. Tveit. Peer-to-peer based recommendations for mobile commerce. In *WMC '01: Proceedings of the 1st international workshop on Mobile commerce*, pages 26–29, New York, NY, USA, 2001. ACM.

[19] J. Yan, N. Liu, G. Wang, W. Zhang, Y. Jiang, and Z. Chen. How much can behavioral targeting help online advertising? In *Proceedings of the 18th international conference on World wide web*, WWW '09, pages 261–270, New York, NY, USA, 2009. ACM.

[20] C.-N. Ziegler. *Towards Decentralized Recommender Systems.* Verlag Dr. Müller, Saarbrücken, Germany, May 2008.