For sale : Your Data
By : You

Christopher Riederer
Columbia Univ.
New York, USA
mani@cs.columbia.edu

Vijay Erramilli
Telefonica Research
Barcelona, Spain
vijay@tid.es

Augustin Chaintreau
Columbia Univ.
New York, USA
augustin@cs.columbia.edu

Balachander Krishnamurthy
AT&T Labs–Research
NJ, USA
bala@research.att.com

Pablo Rodriguez
Telefonica Research
Barcelona, Spain
pablorr@tid.es

ABSTRACT

Monetizing personal information is a key economic driver of online industry. End-users are becoming more concerned about their privacy, as evidenced by increased media attention. This paper proposes a mechanism called 'transactional' privacy that can be applied to personal information of users. Users decide what personal information about themselves is released and put on *sale* while receiving compensation for it. Aggregators purchase *access* to exploit this information when serving ads to a user. Truthfulness and efficiency, attained through an unlimited supply auction, ensure that the interests of all parties in this transaction are aligned. We demonstrate the effectiveness of transactional privacy for web-browsing using a large mobile trace from a major European capital. We integrate transactional privacy in a privacy-preserving system that curbs leakage of information. These mechanisms combine to form a market of personal information that can be managed by a trusted third party.

1. INTRODUCTION

Online services are largely fueled by the collection and exploitation of personally identifiable information (PII [1]). Online entities collect PII of users in exchange for services and these entities monetize this data primarily via advertisements. Information aggregators[2] have found new and creative ways to collect and exploit this data[3] and are increasingly collecting information outside the scope of their application (DoubleClick, Facebook Connect etc.). Various leakages of PII have been identified in traditional Online Social Networks [10] and their mobile counterparts [11]. As aggregators move into monetizing more of this PII, they end up crossing the 'creepiness' line[4], antagonizing end-users and opening the door to a barrage of bad press[5] and potential legislation[6]. There has also been a rise in firms trading in personal information, for instance RapLeaf profiles end-users with personal names and trades this information, without consent of end-users or compensating them[7].

Proposed solutions to preserve privacy have failed to be adopted [2, 1] while the situation has worsened on the ground with leakage being compounded with linkage [9]. Attempts to provide strong privacy guarantees (e.g., differential privacy) undermine the utility of the aggregators, enforcing constraints. We show that these constraints are unnecessary and an economic rethink can lead to a simple alternative solution.

Solution overview: We propose a mechanism called Transactional Privacy (TP) that enables release of portions of PII by end-users (on a strictly *opt-in* basis) to information aggregators for adequate monetary compensation. We define privacy here in terms of control of *flow and usage* of information and TP helps users decide *what* and *how much* information the aggregators should obtain. While TP is designed to be general enough to

[1]Information which can be used to distinguish or trace an individual's identity either alone or when combined with other information that is linkable to a specific individual

[2]We refer to 3rd party aggregators like DoubleClick and online applications like Google, Facebook, Groupon etc., collectively as aggregators

[3]Scrapers dig deep, http://goo.gl/QwJdJ

[4]Google gets close to the creepy line, http://goo.gl/DWXB8

[5]Facebook in privacy breach, http://goo.gl/42frH, Many android apps leak user privacy data http://goo.gl/qVhlz

[6]Whitehouse to push privacy bill, http://goo.gl/pKamG

[7]http://goo.gl/0s591

Hotnets '11, November 14–15, 2011, Cambridge, MA, USA.

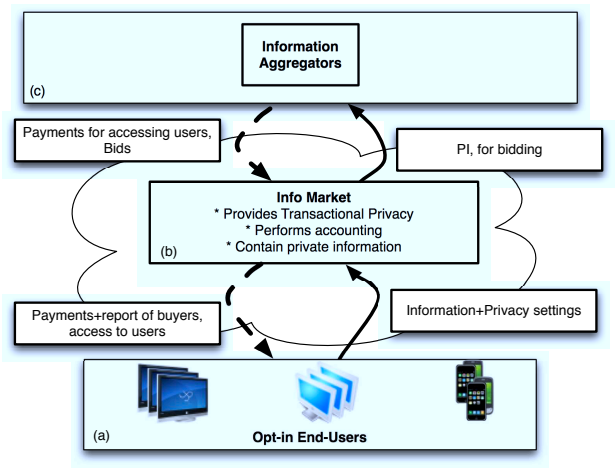


Figure 1: Overview of PIM (a) End-users, (b) Third-party operated Information Market, (c) Information Aggregators

handle different types of PII, such as demographic information, we focus on Web browsing data and location information here. To sell PII, we rely on auctions, where users put up PII and aggregators place bids to gain access to the corresponding user's information. Aggregators can valuate users' PII and decide on the amount to bid, and if they win, gain access to the user with this information for a *limited* time. An important part of our system is to ensure that aggregators cannot strategically manipulate the market and that users are compensated in proportion to aggregators' valuation. Here we leverage prior work on unlimited supply auctions, and in particular the exponential mechanism [13] that is simple to implement and provides good guarantees on truthfulness and market efficiency (see Sec. 2).

We present a simple case study to show how TP can apply to Web browsing behavior, shown to enable leakage of PII [8]. As examples of aggregators, we consider online coupon providers that aim to target users with personalized deals. Using real browsing data from a large number of mobile users and from online coupon deals we study how the revenue of a user changes as a function of the amount of information she releases. We find that, releasing little information initially leads to large increases in potential revenue, while releasing more (potentially sensitive) information yields only a marginal increase in revenue. (see Sec. 3)

We show how TP can be efficiently implemented between aggregators and end-users with a trusted third party in a Personal Information Market (see Fig. 1) that works as follows: (i) an end-user opts-in to the system and decides on information she is willing to disclose about herself; (ii) the third party runs an auction where aggregators can bid to access the user's information through TP; (iii) the third party compensates the end-user (via money or rebates) based on the money exchanged on the market and reports to the end-user about aggregators that received her information, improving transparency; (iv) aggregators who win the auction get access to the user's information associated with the PII they bid on. The system implementing PIM also includes an identity preservation mechanism based on a hybrid browser/proxy architecture that enables such transactions. This mechanism curtails the flow of information to aggregators, protecting against well-known forms of privacy leakages [8, 7], handing back control of PII to the respective end-user. Additionally, as we are proposing an economic transaction, for fair valuation of the information the leakage has to be curbed, forcing aggregators to come to the market (See Sec. 4).

Sec. 5 summarizes benefits of deploying such a system to users, aggregators and other application developers. We then discus related work and the potential research challenges posed by TP.

2. TRANSACTIONAL PRIVACY

Principles

Transactional privacy is guided by three principles: (i) users should have control of their PII and decide what gets released, (ii) aggregators should be able to derive maximum utility of the data they obtain, and (iii) aggregators are best positioned to price the value of users' PII.

Previous work has focused on paying end-users to compensate for their loss of utility [4] via information release, such as the increased chance of being refused insurance if health data is released. The difficult task of calculating the loss of utility was left to the user [2]. An easier and more intuitive task is allowing the user to decide what information she would like released, instead of the utility of that information, while providing relevant information as a *guideline* to aid the user in their decision-making. We focus on web-browsing behavior in this paper, as represented by the set of web-sites a user visits. Note that detailed information about each visit (time spent on a site, etc.) can be easily incorporated. We propose to show the user (via a simple browser plug-in, Sec. 4) the set of sites she has visited in a sorted order (descending) according to their global *popularity*—the number of other users who have visited that site. The first site will be the most visited site etc. User Alice who releases a site with high global popularity (say `facebook.com`) has lower risk of being identified as compared to user Bob who chooses to release `rarecomics.com`, a niche site (tracking notion of k-anonymity [17]).

Providing the aggregator access to *raw* information is in contrast to previous solutions that constrain the aggregators to access data through limited variables that are deemed 'safe' to release [4]. Many aggregators run specialized algorithms on their data sets. Forcing ag-

gregators to disclose these algorithms or constraining the data they are able to use is a losing proposition.

Here is why we believe that aggregators can compute the value of access to a user accurately: First, aggregators have experience extracting value from PII. Second, they are able to assess revenues on a short-term basis through the sale of goods or ad-space, compared to the long-term risk a user must calculate in dealing with privacy. Finally, aggregators typically deal with many customers, and can take a little more risk in overestimating or underestimating the value of access, as opposed to users who are more risk averse.

Model

Formally, we denote the set of users by $\mathcal{I}$, and each user by the index i. The scheme we describe next is general enough to apply to different types of PII. We introduce the set of *sites* $\mathcal{J}$ whose elements, denoted by the index j can be either a URL (for web-browsing), or a geographical location (*e.g.*, a longitude and latitude using GPS, or a cell in a mobile network). We assume that users disclose a simple count of their activity on different sites, denoted by $\mu_i(j)$, which is a vector that indicates how many visits the user has made to either a URL or a location. It is possible to apply the same model to a more complex vector that would indicate time, duration, or order of visits. We assume that each user indicates a subset $S_i \subseteq \mathcal{J}$ that contains all the sites she is ready to be tracked on. This indicates that an aggregator would be able to uniquely identify this user whenever she visits these sites, and will also be given $\mu_i(j)$ for $j \in S_i$. This enables the aggregator to build-up a profile over time, to further help with targeting.

Let us denote the set of aggregators by $\mathcal{K}$, each indexed by k. Intuitively, aggregator k should be willing to pay to access this information as long as the price to acquire it is smaller than the additional revenue r_k it can make. Note that the good being sold on the market is access to PII. This good can be sold to multiple aggregators with no marginal cost of reproduction, hence the market can be thought of as having an unlimited supply. Extensions for an aggregator to buy exclusive access can be included although beyond the scope of this paper. However, there can be strong incentive for aggregators to lie about their valuation.

In order to effectively trade such unlimited supply goods, we rely on the auction mechanism called the exponential mechanism [13] which has the following properties: (i) it has been shown to be a truth telling mechanism; it is in the best interest of the bidders to be honest about their valuation and (ii) the scheme has been shown to be close to optimal in terms of revenue for the seller (end-user in our case). We choose this objective for this paper, while noting that other objective functions (*e.g.*, maximizing revenue for all players in the value chain) can be chosen.

In the auction, we assume that each aggregator k in $\mathcal{K}$ bids a maximum price $p_{i,k}$ that it is ready to pay to access user i. Assuming that the fixed price set is p and all willing bidders pay p, the total revenue is given by:

$$R\left((p_{i,k})_{k\in\mathcal{K}}, p\right) = \sum_{k\in\mathcal{K}} p \times \mathbb{I}_{\{p\leq p_{i,k}\}} \ .$$

When $p > \max_{k\in\mathcal{K}} p_{i,k}$, the revenue will be zero, as no one buys the information that is priced too high.

We wish to choose p to maximize this sum. Following [13] we first assign an initial value to p according to a measure ν on $\mathbb{R}$ and then we re-weigh this measure to choose the actual price used. To re-weigh, we use an exponential function that puts more weight on high value of R, according to a parameter $\varepsilon > 0$. Hence the pdf of the chosen price is given by

$$\frac{\exp\left(\varepsilon R\left((p_{i,k})_{k\in\mathcal{K}}, p\right)\right)\nu(p)}{\int_0^\infty \exp\left(\varepsilon R\left((p_{i,k})_{k\in\mathcal{K}}, s\right)\right)\nu(s)ds}$$

Note that this density is always defined as long as the integral is finite, and note that the function R is zero for p sufficiently large. A natural and simple choice is then to choose the initial distribution of p according to the Lebesgue measure on $\mathbb{R}$, such that $\nu(p) = 1$.

By using ε, we have added noise around the value maximizing the revenue, given the set of bids. Although it seems counter-intuitive to use a suboptimal price, it is shown [13] that this (1) prevents any bidder from winning more than a factor $\exp(\varepsilon)$ when cheating and (2) still reaches a revenue that is within a good bound of the optimal value, denoted OPT, if the number of aggregators is large. The expected revenue is at least $OPT - 3\frac{\ln(e+OPT\epsilon^2 m)}{\epsilon}$, where m is the number of buyers in the optimal case. Thus, although the randomization causes revenue from a given set of bids to be lower, truthful bidding means the set of bids will be higher, ending up with better revenue than if we allowed bidders to cheat.

3. CASE STUDY

We next focus our attention on studying how the revenue of a user changes with varying amounts of information release via TP. For this, we rely on real data consisting of an entire day of browsing behavior on mobile phones of several hundred thousand users from a large European capital, collected during the last week of Nov. 2010, by a large provider. While mobile browsing is inherently different from fixed browsing behavior, we believe the size and the scope of the dataset forms a representative sample of browsing behavior. A second dataset obtained from FourSquare gave us similar results, but we omit them for space reasons. We extracted the number of site visits (URLs) and observed a high variance in terms of visits; a long-tail, which has been observed before in related data [5]. The power law

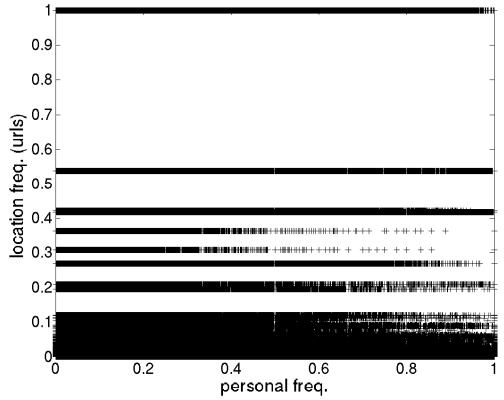


Figure 2: Fraction of time spent by user per site (x-axis) vs. Normalized popularity of sites (y-axis)

fits with exponent 1.5 for mobile browsing passed the Kolmogrov-Smirnov test [3].

For every user, we calculate the fraction of time (in terms of visits) spent on each of the visited sites. For each site she visits, we plot her fraction of time spent on that site versus the global popularity of that site (normalized by the most popular site, `facebook.com`) in Fig. 2. We posit that high values on the x-axis and low values on the y-axis relate to sensitive information. For example, we found that URLs occupying this can be either highly regional, `sarbast.net` or related to a health condition `breastcancer.com`, pertaining to sensitive information [9].

Sample application: Online Coupons

Companies use coupons as a form of price discrimination, that are made more effective with access to PII [14]. Online coupon companies like Groupon have become highly popular and aggregators have shown interests to enter this market[8]. In order to study a user's potential revenue as given by the auction, we use the browsing data and proceed as follows:

(i) For each user, we categorize the URLs of the sites they visited using Alexa.com, which provides the top 500 sites for each category. We filter out visits to ad (*i.e.* Doubleclick, Admob, etc.), analytics, and adult sites to lower any bias.

(ii) We assume that the bidders involved are online coupon vendors and each vendor bids for one category. We found 32 Alexa categories that overlapped with online coupon categories.

(iii) We monitored `yipit.com`, an online coupon aggregator, over three days (July 17-20, 2011) to obtain mean value per deal in each category. We then assume that each user has a likelihood of making a purchase in a category proportional to the fraction of time spent browsing in that category. Thus, the bid values are the mean deal value for a category multiplied by this fraction. The categories `Travel` and `Office Products` had the highest mean values of $844.14 and $207.9.

(iv) For multiple users, we vary the amount of information they reveal. The disclosure strategy is described in Sec. 2, where we release sites in order of popularity from highest to lowest. We release information in blocks of 1% of the volume each time.

(v) For every release, we calculate a set of bids. The majority of high bids came from four `yipit` categories: `computers, home, entertainment, kids_and_teens`.

We pick 4 typical users who have high to middle-level activity and plot (Fig. 3(a)) the optimal revenue they stand to gain as a function of every information release. We obtain the optimal revenue assuming bidders are honest about their valuations. For all of these users, we observe that there is initially a steep increase in revenue with a little disclosure of information, followed by diminishing return as more PII is released. This shows that sensitive information (as given by popularity) is not needed for maximizing revenues. To study enforcement of truth telling in the auction, we plot (Fig. 3(b)) the result of running the auctions for different values of ϵ. Note that smaller values of ϵ enforce truth-telling. We find that the value of ϵ has little or no effect on the results (qualitatively).

4. PERSONAL INFORMATION MARKET

For TP to be effective, we develop a system that curtails the leakage of information and prevents *identification* while browsing. This system should allow users access to all content without being tracked by aggregators while imposing a minimum overhead; we note that it would be impossible to prevent all types of information gathering methods. By raising the bar high enough for information aggregators, we believe they will find it cheaper and more convenient to come to the market.

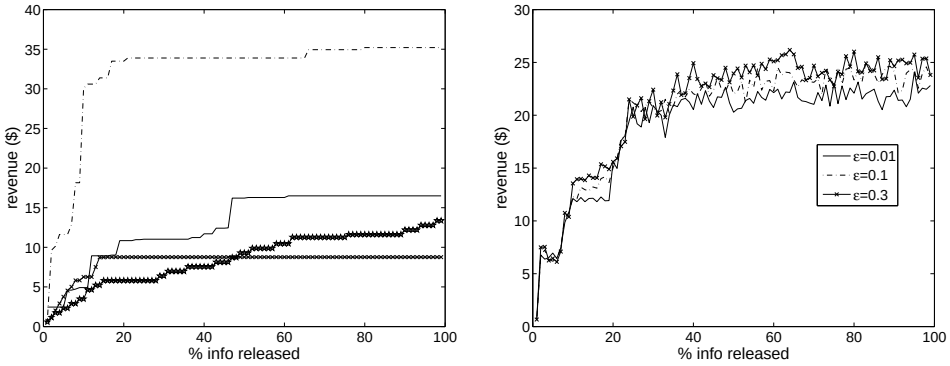
System Description: The full architecture is shown in Fig. 4, with the main additions being a component responsible for transactional privacy and anonymizing proxies in the middle, operated by the trusted third party. At the browser end, a lightweight plugin provides the following functionality: (i) opts-out users of ad-networks and activates Do-not-track[9], showing intent, (ii) provides the user with a mechanism to help them decide which URLs they are willing to put on the market, (iii) prevents leakage (3rd party cookies, super cookies, flash cookies, 1-pixel bugs, etc.) [9], (iv) helps manage multiple users accessing the same device – provides profiles with personalized settings for each user.

For an opt-in user Alice, the operations that take place for Web browsing are as follows:

(i) Alice with IP address IP_{real} browses the web. All her requests go through a proxy. The proxy traps

[8]Facebook jumps into crowded coupon market, http://goo.gl/oLrJy

[9]`http://donottrack.us`



(a) (b)

Figure 3: Revenue for (a) 4 typical users as function of % of information release, (b) Effect of ϵ on the revenue of a user

all `Set-Cookie` HTTP response headers by third parties and masquerades as a legitimate user. No site sees IP_{real} but rather a random IP address (IP_{random}) that changes each time the user visits a new page. This is similar to using a mix-network.

(ii) Alice decides to put her PII up for sale in the auction which can be run regularly (*e.g.*, daily, to near real-time for location).

(iii) If the auction was successful, the anonymizing proxy fixes an IP address (IP_{fixed}) for the user until the next auction is run. IP_{fixed} is passed to the winning bidders, *only* for the sites that were released. Else, the proxy operates as before. In either case, the real IP (IP_{real}) address is never released.

(iv) Suppose that Alice browses to multiple sites belonging to the same aggregator. If the aggregator has purchased this information, it can use this information in any way, such as building a behavioral profile for Alice to entice advertisers. After every auction of Alice's PII, we present IP_{fixed} to the aggregator, so that it can chain multiple purchases.

Note that Alice's future browsing remain monetizable since the IP-fixed can be reassigned. In particular, even if the aggregator accumulates information to profile a user whose information has been purchased, it needs to pay again to recognize this user later.

Online Advertising: Considering online advertising, companies can select targeted ads they want displayed and send them to the aggregator. The aggregator pushes these ads to the user, via the proxy that forwards the ads to the user on the sites she put for sale. If the user Alice clicks on an ad, the anonymizing proxy handles the click, removing the real IP of the user. The proxy establishes a connection to the server hosting the advertisement (can be a CDN or a cloud provider) using the fixed IP (IP_{fixed}) address for Alice so that the advertiser/aggregator can perform accounting. The response is handled by the proxy. Note that even if the advertisers/CDN/cloud provider are in collusion with the aggregator [18], no personal information is leaked (the real IP address is obfuscated).

5. BENEFITS TO PARTIES

End-Users:

End-users choose what to share: The user decides what information is too private and what she is comfortable releasing thus enabling aggregator to avoid negative publicity by only gathering released data.

Monetary elements increase engagement: The good being traded is information, and the end-users are adequately compensated by the sellers. We expect a *positive reinforcing effect* – once users become aware that their information possesses a certain monetary value, they will be more careful with their PII, and privacy increases in the system as a whole.

Opt-in: The system is *opt-in*, which makes end-users valuate their own personal information by themselves - people who are willing to share more information get compensated more and people who want to maintain privacy can do so while forgoing potential rewards.

Aggregators: Information aggregators can use information *without the encumbrance of lawsuits and constant attention* from privacy watchdogs like EFF and privacy researchers. They may access new or better information with the consent of the users, and the nature of the auction ensures they do so whenever it is profitable, preserving utility.

Developers: TP allows application developers to obtain PII for personalized services by *directly* linking them to the owners of the PII: users. We believe this helps developers to decrease capital costs they would incur in building mechanisms to learn more about their respective end-users. In effect we believe that the march towards improving commercial exploitation of personal information can be resolved without crossing the creepiness line with the help of TP.

6. RELATED WORK

Privacy enhancing technologies have evolved over the years. Recent proposals include Adnostic [18] and Privad [6] seek to protect the privacy of the user while enabling aggregators to exploit personal information for targeted advertisements. Such schemes however have

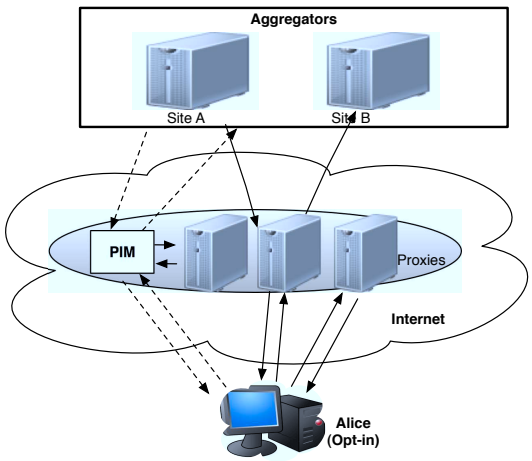


Figure 4: Overview of how browsing can work with PIM and TP

failed to be widely adopted and deployed, for lack of users' incentives [1, 2] and because they undermine the utility for the aggregators by preprocessing data. Running an auction among advertisers while keeping privacy is especially difficult [15]. While transactional privacy has a similar philosophy, we believe that economic incentives for the end-user will increase the adoption and the engagement. We focus on the sale of raw information, albeit with the user's choice and consent. Markets for personal information have been proposed at a high level [12], while current models to compensate users for their privacy [4] make assumptions that are difficult to translate into practice. We have proposed a concrete architecture with transactional privacy at the core to realize such an information market.

7. DISCUSSION

Candidates for trusted third party: The trusted third party has the following roles: act as the legal go-between for the users and the aggregators, implement TP by preventing leakage of users' information, allow users to put information for sale in a transparent manner, run auction mechanisms, enforce payments, and handle any issues from users and aggregators. This can be done for a small percentage of the users' revenues. As we need a party with sufficient resources, we believe the trusted party can be an OS/hardware vendor like Microsoft or Intel, or can be the Telco that provides network connectivity to the user. For the former cases, the ubiquity of the OS and the hardware platform, can make users feel safer as while making them more attractive to aggregators. The advantages for a Telco is that they are highly regulated [16], and users sign a legally binding contract with the Telcos for connectivity that can be extended to cover consent and potential misuse/exploitation of PII. These parties can also control which information is accessed on the device or goes through the network; it may be important to vet both bidders and users to make sure that all provided information is legitimate.

Future work: In this paper we assume that the third-party is trusted; we are currently working on decreasing the level of trust needed. Another model we are considering is to aggregate users into groups before auctioning – increasing the value. An important problem which is an outcome of the presence of the market is the need to prevent *arbitrage*; the ability for aggregators who gained access to sell PII to multiple parties at a cheaper price. Even though we really sell *access* to the user with relevant PII, the longitudinal spread of PII in an illegal way can be detrimental. The other issues involve extending TP to cover different types of PII as well as different types of 'transactions' including transactions conducted with a first-party site.

8. REFERENCES

[1] A. Acquisti. Privacy in electronic commerce and the economics of immediate gratification. In *ACM EC '04*.
[2] A. Acquisti. Nudging privacy: The behavioral economics of personal information. *IEEE Security and Privacy*, 7:82–85, 2009.
[3] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Reviews*, 2007.
[4] A. Ghosh and A. Roth. Selling privacy at auction. In *ACM EC*, 2011.
[5] S. Goel, A. Broder, E. Gabrilovich, and B. Pang. Anatomy of the long tail: ordinary people with extraordinary tastes. In *WSDM*, 2010.
[6] S. Guha, B. Cheng, and P. Francis. Privad: Practical Privacy in Online Advertising. In *NSDI*, 2011.
[7] B. Krishnamurthy. I know what you will do next summer. *ACM CCR*, Oct, 2010.
[8] B. Krishnamurthy, D. Malandrino, and C. E. Wills. Measuring privacy loss and the impact of privacy protection in web browsing. In *ACM SOUPS '07*.
[9] B. Krishnamurthy, K. Naryshkin, and W. Craig. Privacy leakage vs. protection measures: the growing disconnect. *W2SP*, May,2011.
[10] B. Krishnamurthy and C. Wills. On the leakage of personally identifiable information via online social networks. In *ACM WOSN*, August 2009.
[11] B. Krishnamurthy and C. Wills. Privacy leakage in mobile online social networks. In *ACM WOSN*, June 2010.
[12] K. C. Laudon. Markets and privacy. In *ICIS*, 1993.
[13] F. McSherry and K. Talwar. Mechanism design via differential privacy. *IEEE FOCS*, 2007.
[14] A. Odlyzko. Privacy, economics, and price discrimination on the internet. In *ACM ICEC '03*.
[15] A. Reznichenko, S. Guha, and P. Francis. Auctions in Do-Not-Track Compliant Internet Advertising. *ACM CCS*, 2011.
[16] D. C. Sicker, P. Ohm, and D. Grunwald. Legal issues surrounding monitoring during network research. In *ACM IMC '07*.
[17] L. Sweeney. k-anonymity: a model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, Oct. 2002.
[18] Vincent Toubiana, et al. Adnostic: Privacy preserving targeted advertising. In *NDSS*, 2009.