Towards A Privacy Preserving Policy Based Infrastructure for Social Data Access To Enable Scientific Research

Palanivel Kodeswaran*
University of Maryland, Baltimore County
1000 Hilltop Circle
Baltimore, MD 21250
palanik1@cs.umbc.edu

Evelyne Viegas
Microsoft Research
One Microsoft Way
Redmond, WA 98052
evelynev@microsoft.com

***Abstract*— In this paper, we present a policy based infrastructure for social data access with the goal of enabling scientific research, while preserving privacy. We describe motivating application scenarios that could be enabled with the growing number of user datasets such as social networks and medical datasets. These datasets contain sensitive user information and sufficient caution must be exercised while sharing them with third parties to prevent privacy leaks. One of the goals of our framework is to allow users to control how their data is used, while at the same time enable researchers to use the aggregate data for scientific research. We extend existing access control languages to explicitly model user intent in data sharing as well as supporting additional access modes viz. Complete Access, Abstract Access and Statistical Access that go beyond the traditional allow/deny binary semantics of access control. We then describe our policy infrastructure and show how it can be used to enable the above scenarios while still guaranteeing individual privacy. We then present our initial implementation of the framework extending the SecPAL authorization language to account for new roles and operations.**

Privacy; Policy; Social Networks

I. INTRODUCTION

There are an increasing number of users participating in social networks such as facebook [6], where users share personal information with their friends. Similarly, there is an emergence of social networks for other types of data as well such as Covester [2] for finance data and HealthVault for Medical data. However, large amounts of these social data are currently held behind the vaults of large corporations due to legal requirements and privacy considerations of users. On the other hand, users join these networks to share information with their friends as well as benefit from the collective knowledge available in the data set. For example, users enrolled in a medical dataset may benefit from knowing the onslaught of an epidemic in their neighborhood. Similarly, users in a financial dataset may want to see how users sharing a similar portfolio have been doing in the stock market. These queries, although accessing private data represent the aggregate information of a group and are not necessarily privacy revealing. Users may also want to share information of different granularity with their friends depending on the purpose. For example, a user may want to share her zipcode with her friends for mobile social networking applications whereas she may want to share her accurate location for emergency applications. Similarly, researchers may need access to social data to perform research on user trends and network properties. Privacy preserving analysis techniques such as Differential Privacy [4] have been shown to support these kinds of queries without threatening user privacy while enabling valid scientific research [5]. Given the incentive to users for sharing data, there is a need for a framework in which users can specify who can access their data and for what purposes. Unlike traditional access control systems, users may also want the flexibility to control the amount of information that is released based on the purpose for data access. To enable collaborations of such a nature, we propose a policy based infrastructure that allows

1) Users to express their privacy preferences with respect to who can access their data and for what purposes.

2) Data provider support to enforce user privacy preferences as well as supporting additional access modes to release data at different granularities based on the intended purpose.

The main contributions of our work can be summarized as

1) Proposing a policy based infrastructure for sharing social data that is predicated on purpose as well as user identities and attributes of the users.

2) Proposing additional access modes for releasing data at different granularities.

3) Extending traditional access control models to go beyond the binary semantics of allow/deny.

The rest of the paper is organized as follows. We review related work in section II. In sections III and IV, we motivate the case for a formal privacy framework and the need to go beyond binary access control semantics for social networks. We describe sticky policies in section V. We then describe our purpose based access control model and additional access modes in sections VI and VII respectively. We describe our system architecture in section VIII and present our evaluation in section IX. We discuss limitations

*Work performed during summer internship at MSR)

978-1-4244-7556-6/10/$26.00 ©2010 IEEE

and future research in section X and finally conclude in section XI.

II. RELATED WORK

In [1], the authors propose a Data purpose Algebra for computing the acceptable uses of a data as it is transferred among multiple organizations. The allowed set of operations depends not only on the contents of the data, but on the provenance of the data as well. The authors claim that most data transformations and associated purposes can be modeled as algebraic expressions that can later be verified to check if any policy violations were made. Our approach on the other hand is preventive and aims at allowing users to express and enforce their privacy preferences. Carminati et al. [12] propose using semantic web languages for enforcing user privacy in social networks based on various notions of trust relationships such as "friend" and "close friend ". Carminati et al [13] present a rule based approach to access control in which authorization is specified in terms of the type, depth and trust of the relationships existing among users and resources in the network. Ali et al. propose a trust based approach for social data access in which access control is based on the trust relationship between nodes in the network. Persona [10] uses Attribute Based Encryption to enforce user defined access control over data where as Lockr [11] uses attestations of social relationships among users to enforce user privacy. The above approaches address the case of the social network provider not being trustworthy and do not support access to aggregate data. In our approach, the data provider is a trusted entity whose business model depends on satisfying its users and hence protects their privacy. For these data providers, it is desirable to be able to support additional access modes which enhance its users' privacy.

A number of privacy languages viz. P3P [15], EPAL [16], XACML [17] etc. have been developed for expressing privacy policies. However, all these languages possess a binary semantics and do not control the amount of information that is released when access is granted.

III. TOWARDS A FORMAL POLICY FRAMEWORK FOR SOCIAL NETWORKS

In this section we motivate the need for a formal model to make principled decisions about access control with the goal of preserving privacy of individuals in social networks. There has been a growth of decentralized information sources such as social networks in the recent past. In these decentralized settings, the creation and dissemination of data, particularly personal data, cannot be guaranteed for authenticity and privacy. Although most social networks provide basic support for privacy control, these are ad hoc solutions that are hardly understood and used by typical users. For example, facebook allows users to restrict sharing of their profile information and photos to only friends. Furthermore, with the free availability of images on the web, it is easy to download pictures and create duplicate profiles of people in most social networks. There are no inherent tests for authenticity in these networks. Consequently, any inferences derived based on the communities and posts made by the duplicate profile, may not only be incorrect but also be damaging to the person in real life. Therefore, there is clearly a need for users to be able to specify which pieces of their personal information are used by whom and for what purposes.

For the rest of the paper, we consider the following running example. Alice travels with her friends Bob and Cathy and their friend Will on a vacation. When they return, Will uploads pictures of the trip on his favorite social network. However, Alice has a privacy preference that only her friends should be able to identify her in pictures, for example, through the use of tags. Note this is distinct from the case where users have access to pictures containing Alice, possibly due to her shared presence with others. Therefore, under these settings, Will and his friends can still access pictures containing both Alice and Will, although Alice and Will may not be friends. In fact, Will's privacy policy decides who can access the trip pictures containing Alice. In the worst case, everyone can access the trip pictures containing Alice. In this example, we focus on who can tag Alice in pictures. Further, assume that Will has no privacy preferences and allows everyone to access his pictures. Now when Will tags Alice in a picture, all of Will's friends can access and identify her in the picture violating her privacy requirements. This is distinct from the case where Alice denies access to her pictures for non-friends. In this case, Alice's shared presence with Will reveals her privacy. Clearly, there is a need for a formal framework that can jointly reason about both Alice's and Will's privacy policies and evaluate which pieces of data can be shared with whom and for what purposes and what operations are allowed on them. We formalize the above discussion in prolog like syntax below. Consider the following policy of the social network provider. In the rest of the paper, we use social network and data provider interchangeably.

```
hasAccess(X,P) :- owner(Y,P) ^
          sharesWith(Y,P,X)          (1)
```

This rule states that a user can access a picture P as long as the owner shares it with the user.

```
visibleIn(X,P) :- tag(Y,X,P) ^ Person(Y)
                                     (2)
```

The above rule states that X is visible/identifiable in Picture P as long as any user tags X in P.

```
visibleTo(X,Y) :- visibleIn(X,P) ^
                  hasAccess(Y,P)     (3)
```

The above rule is interpreted as X being visible to Y, if Y has access to a picture in which X is identifiable.

Now consider Alice's privacy preferences. Alice wants pictures owned by her to be accessible only to her friends. Furthermore, Alice wants to be visible/identifiable only to her friends. The following rules represent Alice's preferences

```
sharesWith(Alice,P,X) :- friend(Alice,X)
```

```
visibleTo(Alice,Y) :- friend(Alice,Y)
```

Will on the other hand allows all users to access and tag his pictures. This is represented as follows

```
sharesWith(Will,P,X) :- user(X)
```

Now consider the case of whether Will can be allowed to tag Alice in their common picture. Assuming Will is allowed to tag Alice in picture P, we have the following inference

`visibleIn(Alice,P)` from (2)

Now consider user Dave who is not a friend of Alice. Dave has access to P since Will allows access to all. Further, we can make the following inferences

`hasAccess(Dave,P)` from (1)

`visibleTo(Alice,Dave)` from (3)

This is a violation of Alice's privacy preferences since Dave who is not Alice's friend can still identify Alice. Therefore the privacy framework should not allow Will to tag Alice in the picture. Note that, it is Will's privacy policy or lack of it, that threatens Alice's privacy in the shared picture. We would further like to note that restricting tagging to Alice's friends alone would not satisfy her privacy preferences as explained below. Consider the scenario where Bob, who is Alice's friend, tags her in a picture. Unless Bob has a privacy policy that restricts access to their common friends, Alice's privacy preferences would still remain violated. Therefore, in shared settings such as the above, it is necessary to reason about the privacy policies of all participants before sharing data. On the other hand, access to individual data such as personal profiles could still be controlled only by the privacy policy of the concerned individual. In addition to individual users' privacy policies, the social network provider may have their own privacy policies. These policies may arise for various reasons such as societal and legal requirements. For example, the social network provider may not want to allow users to share pirated videos even with their friends. Under these scenarios, the privacy policy of the social network/ data provider must also be considered before making access control decisions.

IV. THE CASE FOR GOING BEYOND THE BINARY SEMANTICS OF ACCESS CONTROL

In this section we motivate the case to go beyond the traditional binary semantics of access control for social data.

Traditional access control models exhibit binary semantics in that a user is either granted or denied access to a resource. These models were typically developed for access control in operating and distributed systems. For example, access control in the linux file systems is achieved through the read, write and execute bits. Similarly, obtaining access to a resource in a distributed system generally requires the requester to present an authorization credential to the system. Furthermore, when access is granted, the user is provided complete access to the resource. For example, when a user is granted read access to a file, the user has access to read all the information in the file. However, this binary semantics may not be applicable to social network data in general. Social networks represent a paradigm shift with respect to sharing of data. Most users join social networks with the main aim of being able to share information with their friends. Unlike traditional systems, the primary goal in social networks is to share. And there is a huge incentive for users to share information with each other in the network and benefit from their collective knowledge. For example, by sharing music tastes, users could get song recommendations that are based on the song history of similar users. Another useful application is locating traffic jams based on traffic data shared by users. In the context of medical social networks, researchers may be able to detect epidemics earlier based on medical data shared by users in an area. In general, there are a large number of useful applications that could exploit the shared data and directly benefit the users. However, most of these applications do not require complete access to the data. Emerging privacy preserving data analysis techniques could be used in these applications to provide the same end benefit to the users. In these scenarios, traditional access control models would not suffice. Consider a traffic application that requires user's GPS co-ordinates to estimate traffic. In the traditional model, the application is either granted or denied access to the actual GPS data depending on the user's location privacy policy. However, using data perturbation techniques, we could add noise to the GPS data and share only the noisy version. In this case, we can still obtain a fairly accurate traffic estimate as well satisfy the user's privacy policy. Traditional access control models are not equipped to handle the latter scenarios in that they cannot control the amount of information that is released once access is granted. The above scenario motivates the need for a framework that allows users to specify their privacy policy in terms of who can access which pieces of their data, for what purposes and how. Making access control decisions based on the three principals of who, what and why provides users with much more flexibility in terms of granting access as well as deciding the representation of the data to share. For example, a user may specify that her date of birth can be used for demographic purposes but should not be used for targeted advertising. Similarly, a user may specify that while sharing her GPS data for emergency situations, no noise be added. On the other hand, when her GPS data is shared with a traffic application, only a noisy version should be shared. In this way, the user can not only specify the conditions for access, but also control the amount of information that is released. Similar to user preferences, the data provider may also have their own policies with respect to how the aggregate dataset is used. In particular, to enable scientific research, data providers may provide researchers access to the dataset, but constrain them to only use predefined privacy preserving data analysis techniques that return only aggregates.

V. STICKY POLICIES

One of the main goals of our framework is to ensure that users have control over how their data is used. Sticky policies which can be viewed as being tied to a piece of data represent the acceptable uses of the data. Consequently, sticky policies can be used to govern access to protected data. While most of

the existing access control policies specify who can access data, our policies also include specification of the purpose for which access is to be allowed. As seen in Fig 1, the sticky policy specifies that "Phone Number" can be used for the purposes of emergency contact where as it is not acceptable to use phone number for marketing purposes.

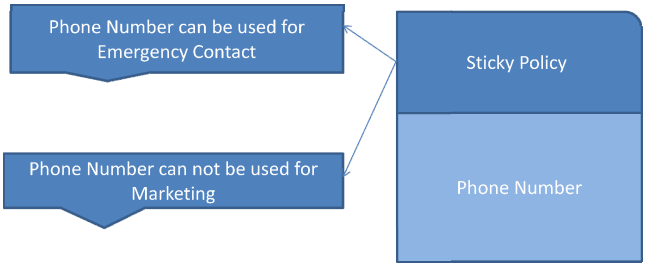

Figure 1. Sticky Policy

Sticky policies could apply to entire datasets as well in addition to individual records. In these cases, the appropriate sticky policy should govern access depending on the data being requested. For example, the policy applied to an anonymized dataset may be completely different from the policy specified by an individual for her information in the original non-anonymized dataset. Ideally, we would like the combined data set policy to release only as much private information as deemed acceptable by the individual's privacy policy.

VI. PURPOSE BASED ACCESS CONTROL

Our framework allows users to specify their preferences with regards to what data to share with whom and for what purposes. In our framework, we explicitly consider the purpose for which data is requested. Based on user specified preferences, the system decides whether the requestor has access to the data and if yes, the mode of access. Our framework supports multiple access modes that can be specified by the user for each authorized access. We now describe the user and data provider preferences supported in our framework.

1) User Preferences

User preferences can be in terms of identities and attributes. The attributes could be in terms of attributes of the user or attributes of the data. User attributes could include the relationships the user has with other users such as being someone's doctor, spouse and so on. Data attributes apply to the data that is being requested and could include the category of the data such as private or public as well as purpose of the data such as emergency contact or public address. Users can then express policies of the form "My Doctor can access my emergency contact number".

2) Data provider Policy

Similar to user preferences, data providers also have a privacy policy with regards to how their dataset is used. Such preferences arise from a variety of reasons such as privacy laws, contractual agreements with the user and so on. For example, to protect user privacy, the data provider may allow researchers to only access the aggregate data and never allow individual data items to be released. Furthermore, the data providers need to enforce the sharing preferences of users in their system. In our framework, the data provider is responsible for both enforcing user preferences as well as guaranteeing user privacy while allowing access to the aggregate dataset.

VII. ACCESS MODES

In this section we describe how users and the data provider can work together to enforce user privacy as well as provide access to aggregate data for scientific research. We define three access modes that differ in the granularity of data released

A. Complete Access

This is similar to read access in traditional access control systems. In this case, the requester is provided complete access to the actual data. Access is typically predicated on the trust relationship between the resource owner and requester. The trust relationship must be explicitly specified by the resource owner. For example, a user may specify that her Doctor has Complete Access to her medical record.

B. Abstract Access

This access mode supports releasing a higher level abstraction of data to the requestor. Higher levels of abstraction depend on the type of data and include pie chart representations and city/state level location information. This access mode requires support from the data provider who must implement an appropriate method for releasing an abstract representation of the data. The actual data representation chosen by the data provider depends on the nature of the data that is being shared.

C. Statistical Access

This access mode is designed for researchers to gain statistical access to aggregate data. The underlying implementation should ensure that researchers can perform valid research while ensuring that user privacy is guaranteed. While the above two access modes are used for enforcing user preferences, statistical access is used to enforce the data provider's privacy preferences while allowing researchers access to social data. In our current framework, we choose Differential privacy [4] as the underlying implementation to provide statistical access.

VIII. POLICY BASED FRAMEWORK FOR SOCIAL DATA ACCESS

We propose a policy based infrastructure for data sharing as it possesses a number of advantages. A policy based infrastructure enables the easy specification of access control policies by users. Depending on the expressiveness of the policy language, users will be able to specify authorization policies in terms of relationships, resource types, purposes and other contextual information. This allows the users to intuitively specify their desired authorization rules as opposed to dealing with low level implementation details. A policy based approach also naturally supports evolution in dynamic environments. A user merely needs to update the policy to enforce new authorization rules under changing environments. A policy specification also enables reasoning and could be

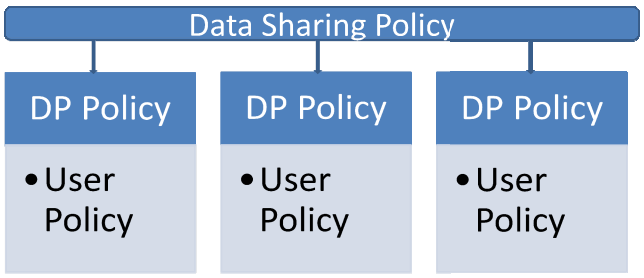

Figure 2. Policy Infrastructure realized through delegation chain

useful in merging and resolving conflicting policies when multiple policies need to be enforced simultaneously. Such situations typically arise when multiple pieces of information could be used to satisfy an information query such as email or phone number for contact information. In these cases additional contextual information such as the purpose of contact could be used to decide between the two pieces of information.

Fig 2. shows the hierarchial policy structure used in our framework, realized through a chain of delegations. The Data sharing policy is a thin layer that arbitrates all access control decisions. The data sharing policy delegates access control decisions to the respective data providers. The data providers in turn delegate access control to User policies. Complete and Statistical Access requests are allowed as long as the user permits it, while statistical access requests are permitted as long as the data provider permits such an access. In this way, the data sharing policy enforces both data provider as well as user privacy policies.

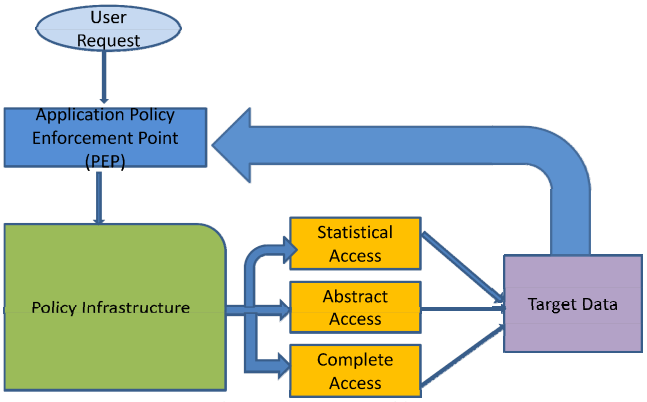

Figure 3. System Architecture

Fig 3. presents our system architecture. At the Application Privacy Enforcement Point (PEP), the user request for data is evaluated by the policy infrastructure along with the access mode. If the request is permitted, the appropriate access mode is applied on the target data and returned to the user.

IX. Evaluation

To evaluate our framework, we verified our approach on a sample dataset that we created from the UCI Census data [7]. We augmented the UCI dataset with a manually generated user id and a salary that is randomly chosen between $0 and $100K.

A. Implementation

We used SecPAL [8] for policy specifications in our framework. SecPAL is a simple yet powerful language that can express most of the commonly used policy idioms. The language has only three deduction rules for

i. Conditional statements
ii. Delegation statements
iii. Can act as statements

These deduction rules completely define the semantics of policies expressed in SecPAL without falling back on other existing logic languages. Since we couldn't express purpose as a first class citizen in SecPAL, we created new Verbs that represent the purpose of data access as follows. Alice's policy "Alice says Bob can AbstractAccess /MyLocation for SocialNetworking" would be represented in our framework as "Alice says Bob can SocialAbstractAccess /MyLocation". We would like to note that this implementation hack stems from our choice of language and later versions of SecPAL such as SecPAL for Privacy [9] have explicit support for expressing purpose and obligations in privacy policies. For the rest of this section, we use "Age" as the running example for data that is sensitive. In our implementation we support the following access modes along with the corresponding output

1) Complete Access : Returns actual age
2) Friendly Access (a form of Abstract access) : Returns an age group such as 30-40
3) Statistical Access

The policy infrastructure itself is set up through delegation chains as follows (The Local Administrator (LA) stands for the authority that finally decides on access control)

LA Delegation to Data Provider

```
i.   LA says %DP canSay %x can read/AbstractAccess/StatisticallyAccess %d if %DP isDataProviderOf %d
ii.  LA says DP canSay DP supportsFriendlyRelease of %d if DP isDataProviderOf %d
```

DP delegation to User

```
i.   MS says %x canSay %y read %d if {%x owns %d, MS isDataProviderOf %d, %x trusts %y}
ii.  MS says %y friendlyAccess %d if {%x owns %d,%x isFriendOf %y,%x allowsFriendlyRelease %d}
```

User Interaction with DP

```
Alice says Alice allowsFriendlyRelease %d if {Alice owns %d, DP supportsFriendlyRelease %d, DP isDataProviderOf %d }
```

Attribute Based Access Control for Statistical Access

```
MS says %y statisticallyAccess %d if{MS
isDataProviderOf %d,%y Possess a,a matches
RoleName="Researcher"}
```

We evaluated our prototype using the following scenario with respect to Alice's data in which Alice trusts Bob and is a friend of Cathy.

```
LA says Bob read /Alice/Age  returns
                 AliceAge=39
LA says Cathy friendlyAccess /Alice/Age
returns          AliceAge = 30-40
```

For statistical access we plot user count against age for different privacy guarantees enforced by differential privacy. Fig 4. shows the result obtained through statistical access for different values of ϵ. The ϵ used in statistical access depends on the trust relationship of the researcher and the fields accessed in the query. Appropriate ϵvalues can be set using approaches similar to those in [5].

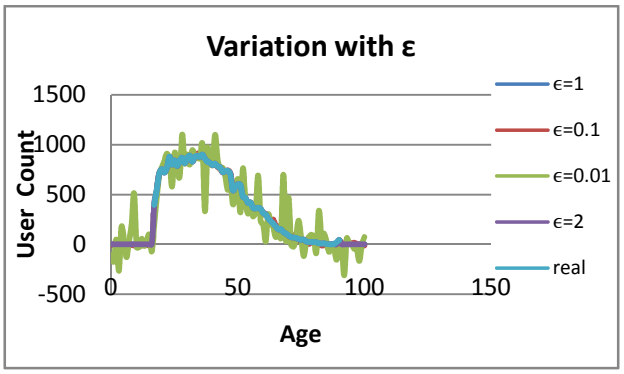

Figure 4. Variation of User count with age for different ϵ

X. Discussion

We would like to note that our framework does not track the usage of data once it is released. Therefore, although our framework allows users to specify which pieces of their personal data can be shared with whom and for what purposes, the framework itself does not guarantee that the data is used by the recipient only for the allowed purposes once it is released. However, we could ensure that the released data is used only for acceptable purposes by using an auditing framework as discussed in [1]. Also, one could fathom scenarios where the user policy may be in conflict with the data provider policy. For example, the user may want to release her date of birth, whereas the data provider policy may have classified this field as private and not permit sharing. In these cases, we pursue a conservative approach and follow the data provider policy. Similarly, there may be situations where the user privacy policy may be weak and hence reveal the privacy of the individual. In these cases, we could extend our framework to include a template of acceptable policies as well as advise users on which pieces of information and combinations thereof could be privacy revealing to help them design a stronger user privacy policy.

XI. Conclusion

In this paper, we have proposed a policy based infrastructure for sharing social data to enable scientific research while preserving user privacy. Our framework allows users to express privacy policies in terms of who can access their data as well as the purpose for which data access is allowed. We extend traditional access control models to go beyond the binary semantics of allow/deny and define new access modes viz. Complete, Abstract and Statistical access that release data at different granularities. Our framework allows Data providers to enforce user privacy policies as well as their own privacy policy while providing researchers access to the data. We have developed our initial framework in SecPAL and verified it on a sample UCI census dataset using scenario based tests.

Acknowledgment

We would like to thank the anonymous reviewers for their valuable reviews and insightful feedback.

References

[1] C. Hanson, T. Berners-Lee, L. Kagal, G.J. Sussman and D. Weitzner. 2007. Data-Purpose Algebra: Modeling Data Usage Policies. In *Proceedings of the Eighth IEEE international Workshop on Policies For Distributed Systems and Networks* (June 13 - 15, 2007). POLICY. IEEE Computer Society, Washington, DC, 173-177. DOI= http://dx.doi.org/10.1109/POLICY.2007.14

[2] Covestor, http://www.covestor.com/

[3] HealthVault, http://www.healthvault.com/

[4] C. Dwork. "Differential privacy". In M. Bugliesi, B. Preneel, V. Sassone, and I. Wegener, editors, ICALP (2), volume 4052 of Lecture Notes in Computer Science, pages 1–12. Springer, 2006.

[5] P. Kodeswaran, and E. Viegas. 2009. Applying differential privacy to search queries in a policy based interactive framework. In *Proceeding* PAVLAD '09.

[6] Facebook, http://www.facebook.com

[7] CenusDataset, http://archive.ics.uci.edu/ml/datasets/Census+Income

[8] M. Y. Becker, C. Fournet, and A. D. Gordon. SecPAL: Design and semantics of a decentralized authorization language. In 20th IEEE Computer Security Foundations Symposium (CSF), pages 3–15.

[9] SecPAL for Privacy, http://research.microsoft.com/apps/pubs/default.aspx?id=10261

[10] R. Baden, A. Bender, N. Spring, B. Bhattacharjee and D. Starin. 2009. Persona: an online social network with user-defined privacy. *SIGCOMM Comput*

[11] A. Tootoonchian, K.K. Gollu, S. Saroiu, Y. Ganjali and A. Wolman 2008. Lockr: social access control for *Web 2.0*. In Proc. WOSN 2008.

[12] B. Carminati, E. Ferrari, R. Heatherly, M. Kantarcioglu and B. Thuraisingham. 2009. A semantic web based framework for social network access control. In Proceedings of the 14th ACM Symposium on Access Control Models and Technologies (Stresa, Italy, June 03 - 05, 2009). SACMAT '09. ACM, New York, NY, 177-186. DOI= http://doi.acm.org/10.1145/1542207.1542237

[13] B. Carminati, E. Ferrari and A. Perego. Enforcing Access Control in Web-based Social Networks. ACM Transactions on Information & System Security, 2008.

[14] B.Ali, W. Villegas and M. Maheswaran. A Trust based Approach to protecting user data in social networks. In 2007, Conference of the Center for Advanced Studies on collaborative Research (CASCON) 2007.

[15] P3P, http://www.w3.org/P3P/

[16] EPAL, http://www.zurich.ibm.com/security/enterprise-privacy/epal/Specification/index.html

[17] XACML, http://*www.oasis-open.org/committees/xacml/*