# A Policy Based Infrastructure for Social Data Access with Privacy Guarantees

Palanivel Kodeswaran*
Dept of CSEE, UMBC
Baltimore, MD 21250
palanik1@cs.umbc.edu

Evelyne Viegas
Microsoft Research
Seattle, WA, USA
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, medical datasets etc. 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 such as Covester [2] for finance data, or HealthVault [3] 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 know how users sharing a similar portfolio have been doing in the stock market. These queries, although they access private data, they 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 zip code 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 or network properties. Privacy preserving analysis techniques such as Differential Privacy [4] have been shown to support these kinds of queries without threatening user privacy and while enabling valid scientific research [5]. To enable collaborations such as those mentioned above, 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.

*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.

## II. RELATED WORK

In [1], the authors propose a Data purpose Algebra for computing the acceptable uses of 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. A number of recent works address privacy challenges in social networks, and we do not intend to provide a complete survey here. Carminati et al.

*Work done during summer internship at MSR

[12] propose a rule based access control using semantic web languages for enforcing user privacy in social networks based on various notions of trust relationships including depth such as "friend", "close friend." Persona [10] uses Attribute Based Encryption to enforce user defined access control over data whereas 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.

## III. STICKY POLICIES

One of the goals of our framework is to ensure that users have control over how their data is used. In this context, sticky policies which can be viewed as being tied to a piece of data can be used to govern access to protected data.

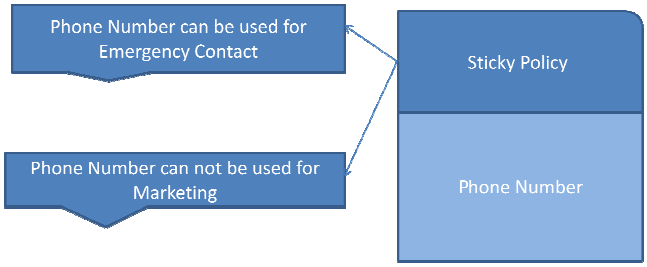


Figure 1. Sticky Policy

In our framework, all accesses to private data need to be authorized by the corresponding sticky policy. 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 purposes of emergency contact whereas it is not acceptable to use phone number for marketing purposes. Sticky policies could apply at different levels of abstraction from individual data to entire datasets. 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.

## IV. 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.

## V. 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 representation of data to the requestor. Higher levels of abstraction include pie chart representations, city/state level location information, etc. 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.

## VI. POLICY BASED FRAMEWORK FOR SOCIAL DATA ACCESS

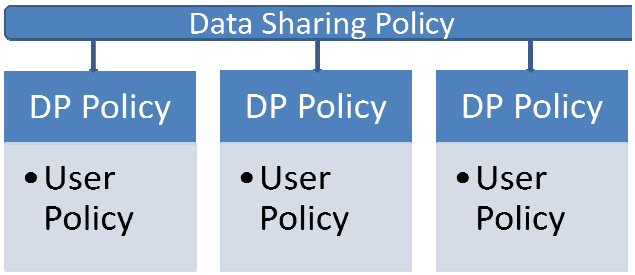


**Figure 2. Policy Infrastructure realized through delegation chain**

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 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 allows such an access. In this way, the data sharing policy enforces both data provider as well as user privacy policies.

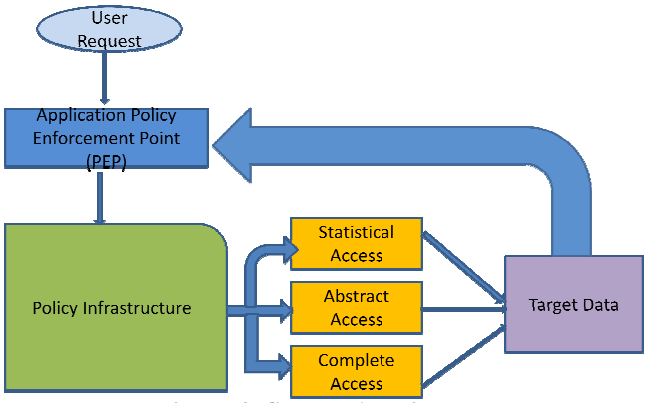


**Figure 3. System Architecture**

Fig 3. presents our system architecture. At the Application Privacy Enforcement Point (PEP), the user's 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.

## VII. 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 to demonstrate different access modes.

### 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 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 (DP)**

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 $\epsilon$. The $\epsilon$ used in statistical access depends on the relationship of the researcher and the fields accessed and can be set using approaches similar to those in [5].

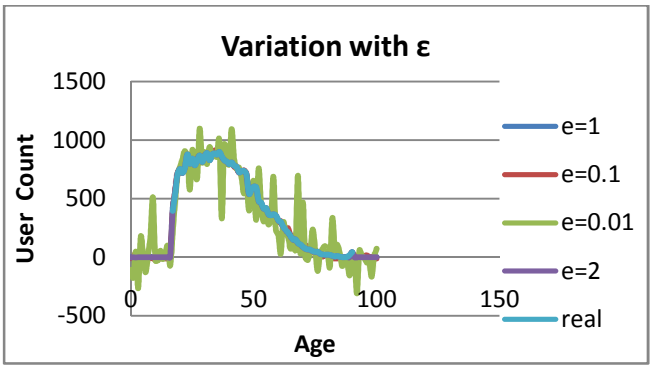


**FIGURE 4. VARIATION OF USER COUNT WITH AGE FOR DIFFERENT $\epsilon$**

## VIII. 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 allowing researchers access to the data. We have developed our framework in SecPAL and verified it on a sample UCI census dataset using scenario based tests.

## REFERENCES

[1] Hanson, C., Berners-Lee, T., Kagal, L., Sussman, G. J., and Weitzner, D. 2007. Data-Purpose Algebra: Modeling Data Usage Policies. In Proc. POLICY 2010. Washington, DC.

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

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

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

[7] 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] http://research.microsoft.com/apps/pubs/default.aspx?id=10261

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

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

[12] Carminati, B., Ferrari, E., Heatherly, R., Kantarcioglu, M., and Thuraisingham, B. 2009. A semantic web based framework for social network access control. In Proceedings SACMAT '09. ACM, New York, NY, 177-186.