GlobalInferencer: Linking Personal Social Content with Data on the Web

Sharon Paradesi
Massachusetts Institute of Technology
paradesi@csail.mit.edu

Fuming Shih
Massachusetts Institute of Technology
fuming@mit.edu

ABSTRACT

The past year has seen a growing public awareness of the privacy risks of social networking through personal information that people voluntarily disclose. A spotlight has accordingly been turned on the disclosure policies of social networking sites and on mechanisms for restricting access to personal information on Facebook and other sites. But this is not sufficient to address privacy concerns in a world where Web-based data mining tools can let anyone infer information about others by combining data from multiple sources. To illustrate this, we are building a demonstration data miner, GlobalInferencer, that makes inferences about an individual's lifestyle and other behavior. GlobalInferencer uses linked data technology to perform unified searches across Facebook, Flickr, and public data sites. It demonstrates that controlling access to personal information on individual social networking sites is not an adequate framework for protecting privacy, or even for supporting valid inferencing. In addition to access restrictions, there must be mechanisms for maintaining the provenance of information combined from multiple sources, for revealing the context within which information is presented, and for respecting the accountability that determines how information should be used.

1. INTRODUCTION

Prior to the advent of social networks, only the big corporations and agencies kept detailed records of the personal information of an individual. Over the last few years, access to personal information has opened up to the general public. Using various data mining tools, it has become easy to infer about a person's behavior, lifestyle, schedule, etc.

There are three instances where gathering a user's personal information can lead to incorrect inferences. They are (i) lack of complete information, (ii) taking user's data out of context, and, (iii) use of innocuous pieces of information in ways the user never expected. To illustrate (i), just because a user has a few photos of unhealthy foods on her profile, it does not mean that she leads an unhealthy lifestyle. She could be working out everyday but she may not post about that on her profile. To illustrate (ii), a user could update her status with a quote about wine by a famous philosopher, but she might have never consumed any alcoholic beverage. To illustrate (iii), based on the information that a user lives in a particular county, an insurance company can point to government documents and assume that because a majority of the residents in that county have had major heart diseases, she is likely to be affected too.

We demonstrate the risk of data mining using multiple sources with a system called *GlobalInferencer*. *GlobalInferencer* mines popular social networks like Facebook and Flickr to obtain users data. It then combines this data with publicly-available information on the Web. Using this method, inferences that were not possible just by looking at a person's profile become apparent. We believe that multi-source inferencing will become more common and the privacy risks of Social Networks must inevitably be viewed in this light.

The rest of the paper is structured as follows. Section 2 describes two problems when mining data from multiple sources: inappropriate inferences and inappropriate use. Section 3 lays out the design and architecture of the *GlobalInferencer* system and illustrates this with a walkthrough of the health scenario. Finally, we discuss steps that can be taken to address the open challenges.

2. EXPOSING PERSONAL INFORMATION IN A MULTIPLE SOURCES WORLD

Personal information on social network might seem innocuous in itself. However, the seemingly harmless information can result in incorrect inferences about the user. We present two major sources of this risk that arises due to this information exposure.

2.1 Inappropriate Inferences

Inferences from data on the Web can be extremely misleading if they ignore the context and provenance (source of information).

According to [8], context is defined as "any information that can be used to characterize the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves."

In social networks, contexts might include the original intent, the original audience addressed, the content's object references, the original activity or practice. Contexts also

suggest a specific social environment in which the content is produced. In general, when a user reads a piece of information, she usually places it within an implicit context in order to interpret it. Without contexts, the consumers of the information have no frame of reference to appropriately evaluate the message conveyed.

It is a common practice for companies to construct user profile by aggregating the content on the Web. One problem of building profile of a user is that the content aggregator has no easy way to reconstruct the context of the retrieved data. And yet information on the Web is often disclosed with respect to context-related norms [11] that govern how information is intended to be interpreted. A crawler that scrapes only targeted content and aggregates and transmits it without related contextual information can easily violate context-related norms.

2.2 Inappropriate Use

According to [17], information accountability means "the use of information should be transparent so it is possible to determine whether a particular use is appropriate under a given set of rules and that the system enables individuals and institutions to be held accountable for misuse."

Accountability differs from access control in that it is fundamentally associated with purpose. Access control typically regards purpose in a narrow way. The subject requests a particular form of access which translates into some set of physical operations which could have a wide range of abstract functions with predictable consequences. Accountability, on the other hand, is associated with using information in a particular social context - it is not simply about reading or writing information, but about what is done with that information over time.

It is easy to use publicly-available Social Web data for any purpose because there is no framework generally available for a user to specify how her data may be used. Thus, a lack of accountability leaves open the opportunity for inappropriate use of personal information. Two techniques that would help move towards accountable systems are (i) access control limited to individuals that the user trusts will not abuse her data, and, (ii) frameworks that detect misuse with means to obtain redress.

3. GLOBALINFERENCER

GlobalInferencer is an information retrieval tool that gathers data from various sources and makes inferences. We start by explaining its architecture and functionality and then walk through the described scenario.

3.1 Architecture

As shown in Figure 1, *GlobalIndentifier* fetches personal contents of a user from different social networks. These contents are transformed into RDF to model the semantic relationships with predefined ontologies such as FOAF [4] and SIOC [2]. The RDF data from each social network is saved locally in the repository as named graphs along with provenance information [15]. In addition to semantic presentations of the personal contents, *GlobalIndentifier* has data miners that simulate how a user's contents can be analyzed by using machine learning or natural language processing techniques. For example, some data miners search for co-occurrences of keywords and infer possible relations among the data. Also, the external knowledge from linked-data sources such as data.gov [9] or DBPedia [1] can be combined to introduce additional background knowledge into the reasoning process. For example, the description of the neighborhood where the user lives can suggest the social status of the user.

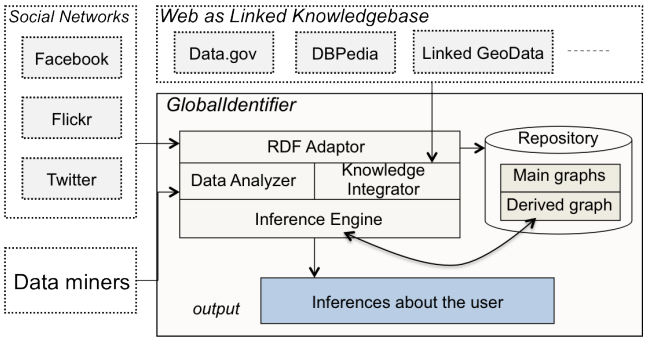

Figure 1: Overview of the architecture of *GlobalInferencer*

3.2 Motivating Scenario

We describe a scenario that highlights a specific case of the challenges mentioned above. We demonstrate how a health-insurance company can use personal information on social networks to corroborate various decisions. In our scenario, an employee of the health-insurance company uses *GlobalInferencer* to learn about an applicant's lifestyle. The data miner searches the Web to see if there are any pieces of information that can be linked to the user's social network to verify whether they are healthy or not. This is explained in detail in Section 4. We expect this to be a growing trend in the future where insurance companies, employers and other agencies will turn to personal information on the Web to make important decisions.

3.3 Scenario Walk-through

The current demonstration version of *GlobalInferencer* provides two canned queries allowing the health insurance company to perform a search on a fictitious user, called Danny. In the first case, the company can learn about the dietary habits of a user, while in the second, it can learn about the user's lifestyle.

The first query is 'Does Danny have an unhealthy diet?'. *GlobalInferencer* tries to first understand what 'unhealthy diet' means by searching the web for occurrences of that phrase. Using the results [1], it comes up with a list of foods that qualify as being unhealthy. It then searches the photos of Danny from Facebook and Flickr to see if any term from the list occurs there. This process has not been fully automated in the system at present. All photos that contain any of the terms in the list of unhealthy foods either in description or comments are collected. Figure 2 shows the query along with the list and photos. These results could easily lead the insurance company to believe that Danny does not have a healthy diet.

The second query is 'Does Danny have a healthy lifestyle?'. This prompts the data miner to first understand what 'lifestyle'

[1]http://www.livestrong.com/article/292968-a-list-of-non-healthy-foods/

Start a new search for information about Danny

Does Danny have an unhealthy diet?

We found this list of unhealthy foods

breakfast sandwich	cheeseburgersfrench fries	fried fish sandwich	taco
hotdog	chocolate nut candy bar	custard-based pie	chocolate chip
soft drink	submarine sandwich	ham	pretzel
hotdog	pickle	breaded shrimp	fried shrimp
potato salad	pecan pie		

Source: http://livestrong.com/article/292968-a-list-of-non-healthy-foods/

Some of these foods were found in Danny's Social Networks as shown in the following photos

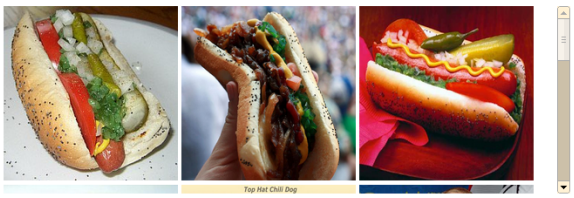

Figure 2: Social Network query results related to diet with additional information from livestrong.com

Start a new search for information about Danny

Does Danny have a healthy lifestyle?

Danny lives in Worcester, MA (source: Facebook). Health statistics for that county is shown below

State Name: Massachusetts, County: Bristol
Percentage of population having health issues:

Health Issue	Population (%)
Blood Pressure	27.7
Diabetes	7
Obesity	23.2
Smoking	24
No Exercise	26.6
Eating only few fruits and vegetables	76.4

source: http://www.data.gov/raw/2159

Furthermore, there were some post(s) from Danny's Facebook showing his dislike of exercise

started to hate exercise :(

Figure 3: Social Network query results related to lifestyle with additional information from data.gov

means. We created a remote repository with data fetched from data.gov [2] to highlight health-related statistics of various counties in the United States. After retrieving Danny's current location (city), *GlobalInferencer* identifies the corresponding county and then obtains the appropriate statistics. Additional evidence of Danny's lifestyle choices comes from Danny's posts that show a dislike towards exercising. Figure 3 shows the query along with the statistics and post collected. These results could lead the insurance company to believe that Danny has an unhealthy lifestyle simply by looking at his current location.

4. FUTURE RESEARCH DIRECTIONS

GlobalInferencer demonstrates the need for the following two enhancements to achieve more reliable inferences.

4.1 Provenance-based Context Construction

[2]http://www.data.gov/raw/2159

While different sources of information can contribute to profiling a user, it is critical to have a *knowledge provenance* system as described in [5]. The provenance component provides the information about the origin of a piece of knowledge, which is critical for agents to establish trust when consuming the data and also weigh how relevant that piece of information is while deriving inferences. We need a context constructor that updates relevant contextual information dynamically during the reasoning process. The context constructor yields additional contextual information to help the data consumer better interpret a user's personal information.

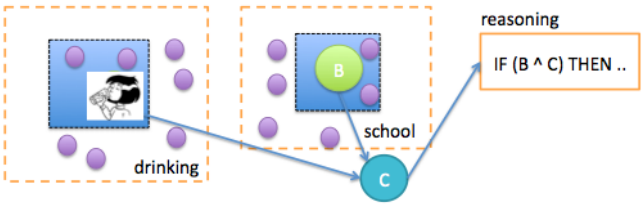

Figure 4: Example of provenance-based context construction

Figure 4 shows an example of provenance-based context construction. Suppose a photo of a user drinking a beverage was uploaded to Facebook, and certain context information such as place and time were captured by the device (depicted within the shaded square). The photo could be interpreted to portray the user consuming alcohol. If this information is further combined with the additional knowledge such as "the user is a teacher" and "the event happened during school hours", the consequences could be rather grave. However, the reality could be that the photo was captured during a demonstration of making cocktail in a cooking class. Contextual information relevant to this fact might not have been observable by the device at the time when the photo was captured (depicted within the dotted square). For example, contextual information such as "the content of the beverage" and "the subject of the class". Note that it is difficult to enforce how data consumer can reason about the contents, but the focus here is to provide a mechanism that regards and updates context information when needed. Thus a user might later add additional information to describe the situation of the event, and the system can either push this new information (automatically or by request) to the entity that parsed the photo previously.

The architecture for preserving context will depend on provenance computations such as those found in scientific workflow systems [6]. Every action on the data should be recorded with provenance information and the system should preserve references for computing relevant contexts. The dynamic and distributed nature of context becomes the main challenge of implementing an efficient algorithm for context re-construction.

4.2 Usage-based Policy

A user should be able to specify how his or her data can be used by an external entity. For example, a photo should not be used for recommending products if it has a policy restrict-

ing its use for commercial purposes. In order to enable this, we need mechanisms to guide the usage of content based on agreed-upon policies as mentioned in [16] and in Social Web data as discussed in [12]. Such systems will ensure accountability in two ways. First, transparency encourages data mining tools to abide by the user's usage policies. Second, it allows data mining tools to justify their use of data and show compliance to user's policies.

5. RELATED WORK

The issues pointed out by [14] are very relevant to this research. Unless the parties that perform data mining of personal information are held accountable to the public, it is difficult to monitor the privacy of the users or ensure that ethical practices are being followed. The paper [14] lists several arguments that could be stated in defense of web-data mining. The second argument states that "There are laws to protect private information. Besides, privacy policies found on many web sites guarantee privacy. So, why worry?". This argument is especially harmful with the proliferation of personal information via social networks. It is unreasonable for one to expect that a user would think in advance of the many ways that a single photo could be used against her. Additionally, linked data gives rise to a new dimension of data mining where one photo or post could be seen in a totally different light than the user originally saw when she posted it on her social network profile.

According to the studies described in [3], users are aware of the implications of their privacy settings and are actively engaged in optimizing their privacy. However, the issue is more than just privacy settings on Facebook. The main thing to be concerned about is the usage of the data even among the entities a user chooses to share her data with (friends, lists, applications and others).

In [7], Debatin et al. hypothesize and later conclude that Facebook users are more likely to perceive risks to other's privacy rather than to their own privacy. The scenarios in this paper deal with basic inferences for an insurance company. However, the possibilities are endless in the future. Usually, by the time a user perceives a risk happening to her, it is usually too late.

As stated in [13], generally known information such as gender, birthday and pictures is more likely to be shared via one's profile, than specific types of information such as phone number or class schedule. In the context of Web data mining, however, this still poses a challenge because general information such as photos are sufficient to prove a case as a prosecutor or, in the case of *GlobalInferencer*, sufficient to build up a strong case of evidence.

In [10], Gross and Acquisti discuss re-identification, which deals with linking non-identifiable datasets (the health statistics in our scenario) with personally identifying datasets (such as Danny's profile). They highlight a privacy implication involving building a digital dossier by continuously monitoring user profiles. Using systems such as *GlobalInferencer*, dossiers can now become global by reaching beyond social networks and into the Web.

6. CONCLUSION

The past year has seen a growing public awareness of the privacy risks of social networking through personal information that people voluntarily disclose. The risk of data mining tools reaching inappropriate inferences and using data inappropriately is more apparent beacause of the ability to search the Web for publicly-available data that are related to a user's personal information. We demonstrate this through a system, called *GlobalInferencer*, that mines popular social networks like Facebook and Flickr to obtain users data. It then combines this personal data with publicly-available information on the Web. It demonstrates that controlling access to personal information on individual social networking sites is not an adequate framework for protecting privacy, or even for supporting valid inferencing. In addition to access restrictions, there must be mechanisms for maintaining the provenance of information combined from multiple sources, for revealing the context within which information is presented, and for respecting the accountability that determines how information should be used.

7. ACKNOWLEDGEMENTS

The authors would like to thank Hal Abelson and Joe Pato for their insightful comments and suggestions. We also appreciate the input received from other members of the Decentralized Information Group at CSAIL, MIT.

8. REFERENCES

[1] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. DBpedia: A nucleus for a web of open data. In *Proceedings of the 6th International Semantic Web Conference (ISWC)*, volume 4825 of *Lecture Notes in Computer Science*, pages 722–735. Springer, 2008.

[2] D. Berrueta, D. Brickley, S. Decker, S. FernÃąndez, C. GÃűrn, A. Harth, T. Heath, K. Idehen, K. Kjernsmo, A. Miles, A. Passant, A. Polleres, L. Polo, and M. Sintek. Sioc core ontology specification. W3c member submission, W3C, June 2007.

[3] D. Boyd and E. Hargittai. Facebook privacy settings: Who cares? *First Monday*, 15(8), 2010.

[4] D. Brickley and L. Miller. FOAF Vocabulary Specification 0.97. Namespace document, January 2010.

[5] P. P. da Silva, D. L. McGuinness, and R. McCool. Knowledge provenance infrastructure. *IEEE Data Eng. Bull.*, 26(4):26–32, 2003.

[6] S. Davidson, S. Cohen-Boulakia, A. Eyal, B. LudÃd'scher, T. McPhillips, S. Bowers, M. K. Anand, and J. Freire. Provenance in scientific workflow systems.

[7] B. Debatin, J. P. Lovejoy, A.-K. Horn, and B. N. Hughes. Facebook and online privacy: Attitudes, behaviors, and unintended consequences. *Journal of Computer-Mediated Communication*, 15(1):83–108, 2009.

[8] A. K. Dey. Understanding and using context. *Personal and Ubiquitous Computing*, 5:4–7, 2001.

[9] L. Ding, D. DiFranzo, A. Graves, J. Michaelis, X. Li, D. L. McGuinness, and J. A. Hendler. Twc data-gov corpus: incrementally generating linked government data from data.gov. In M. Rappa, P. Jones, J. Freire, and S. Chakrabarti, editors, *WWW*, pages 1383–1386. ACM, 2010.

[10] R. Gross and A. Acquisti. Information revelation and privacy in online social networks. In *Proceedings of the 2005 ACM workshop on Privacy in the electronic society*, WPES '05, pages 71–80, New York, NY, USA, 2005. ACM.

[11] H. Nissenbaum. Privacy as contextual integrity. *Wash. L. Rev.*, 79:119, 2004.

[12] F. Shih and S. Paradesi. Saveface: Save GeorgeâĂŹs faces in Social Networks where Contexts Collapse. 2010.

[13] F. Stutzman. An evaluation of identity-sharing behavior in social network communities. *in iDMAa and IMS Code Conference*, 3, 2006.

[14] L. van Wel and L. Royakkers. Ethical issues in web data mining. *Ethics and Information Technology*, 6:129–140, 2004. 10.1023/B:ETIN.0000047476.05912.3d.

[15] E. R. Watkins and D. A. Nicole. Named graphs as a mechanism for reasoning about provenance. 2006.

[16] D. Weitzner, H. Abelson, T. Berners-Lee, C. Hanson, J. Hendler, L. Kagal, D. McGuinness, G. Sussman, and K. Waterman. Transparent accountable data mining: New strategies for privacy protection. *Computer Science and Artificial Intelligence Laboratory Technical Report. www. csail. mit. edu*, 2006.

[17] D. J. Weitzner, H. Abelson, T. Berners-Lee, J. Feigenbaum, J. Hendler, and G. J. Sussman. Information accountability. *Commun. ACM*, 51:82–87, June 2008.