Privacy Preserving IR 2015 - A SIGIR 2015 Workshop

Grace Hui Yang
Georgetown University
huiyang@cs.georgetown.edu

Ian Soboroff
NIST
ian.soboroff@nist.gov

1 Introduction

Information retrieval and information privacy/security are two fast-growing computer science disciplines. There are many synergies and connections between these two disciplines. However, there have been very limited efforts to connect the two. Due to lack of mature techniques in privacy-preserving IR, concerns about privacy and security have become serious obstacles that prevent valuable user data from being used in IR research, for example, query logs, social media, tweets, sessions, and medical record retrieval. This privacy-preserving IR workshop aims to spur research that brings together the research fields of IR and privacy/security, and mitigate privacy threats in information retrieval by constructing novel algorithms and tools.

The schedule of the workshop, papers, presentations, and more can be found on the workshop website at http://privacypreservingir.org/.

2 Keynote: Li Xiong, Emory University

Dr. Li Xiong gave a keynote presentation entitled, "Making Private User Data Accessible for Information Retrieval Research: Data Sharing with Differential Privacy." Dr. Xiong heads the Assured Information Management and Sharing (AIMS) research group at Emory University. This talk introduced the audience to the fundamentals of differential privacy (DP) and its applications in managing user behavior data.

DP allows the synthesis of data D' to closely match the statistical properties of an original dataset D while making strong guarantees that information about individuals in D cannot be recovered from D'. This topic was of great interest to the workshop, as DP is one of the few avenues of work trying to develop provable privacy guarantees for datasets, and the academic IR community is challenged by a lack of access to realistic user behavior data. Dr. Xiong discussed how differential privacy can handle sequential data, which is of high dimensionality and also strongly self-correlated.

This session closed with a wide-ranging discussion of the applications and limits of differential privacy in IR. One proposal was made to start from any of several well-known SIGIR papers featuring log data analysis, and to develop a differentially-private model for the log data underlying

that study. This would simultaneously allow us to understand how to make data differentially private, and also to see what the limits of such data would be with respect to a real study, and lastly make a valuable case to industry.

3 Paper Presentations

The workshop featured five paper presentations. The first presentation was given by Luís Marujo of CMU on privacy-preserving multi-document summarization. The scenario is providing third-party summarization services and ensuring that the privacy concerns of document owners are addressed. The proposed technique has the document owner extract key phrases from the documents using standard NLP tools. These phrases are then obfuscated using secure binary embeddings, and the summarization algorithm works with the embedded phrases rather than the raw text.

Next, Jiyun Luo of Georgetown University presented the work of Sicong Zhang and others on differential privacy applied to search query logs. If companies release query and click data using a differential privacy protocol, we might expect that retrieval effectiveness in a system training and testing from parts of that log to differ from a system that had access to the original log. The experiment in this work measured the effect of using a differentially-private version of a query log with clicks, compared with using the true log.

The third talk was from Simon S. Woo of USC, presenting a short paper on the "security questions" practice used by many websites to enhance security. The sites typically allow the user to select a question from a list of questions, and the user provides an answer that they may be asked to recall at a future authentication time. The problem is that these questions often ask personal information, and often the answers are not difficult to find using web resources. As an end goal, Woo proposes to warn users away from revealing personal information that is findable on the web.

Next, Heng Xu of Pennsylvania State University remotely presented two papers. The first proposed a framework for thinking about mobile app privacy based on Helen Nissenbaum's theory of contextual integrity. The framework identifies context, information attributes, transmission principles, and actors as elements making up privacy. Xu and her team analyzed 49 coloring apps for children looking for violations of transmission principles and what attributes were leaked thereby to how many actors.

Xu's second presentation proposed a framework for providing "nudges" to social network users when they are about to commit an activity that may violate their own or someone else's privacy. For example, sharing a photo where a friend is tagged could violate that friend's privacy. A system might use the friend's sharing behavior as a clue to their privacy preferences; if they don't typically share photos, perhaps they might be concerned about having their photo shared. The system as the authors imagine it would present the sharing user with a "nudge", telling them that they are about to share this photo but that friend does not often share photos, and asking if they are sure they want to do it. This work as well as Woo's inspired a spirited discussion about how systems might make users aware of how personal information is used and when privacy may be at risk.

4 Debate

We hold a debate in the afternoon session. The participants formed two teams, one in favor of and another opposite to the following statement - "Removing Personalization in Information Retrieval would resolve the issue of Privacy in IR". The topic showed a seemingly obvious right or wrong, however, regardless of their own beliefs, the participants were asked to take their assigned side and to find supporting arguments for their own side.

The team that were supporting the statement focused on arguing that privacy issues are mainly caused by using personal profiles and contextual information to help with more targeted retrieval. However, personalized IR is not necessary for IR. Without personalization, an IR system could still perform document retrieval and satisfy the user's information needs. Without personalization, IR systems won't really need personal information from the user to optimize the results according to personal profile. Thus user's privacy could be protected since personal information is not used. Privacy and IR evaluation was also mentioned.

The team holding the opposite argument focused on showing counter examples of how privacy is violated in ad-hoc retrieval without being personalized. Many queries themselves contain personal or sensitive information. It is actually difficult to isolate personalization from the general document retrieval task. The interesting thing was that even it seemed that the opposite arguments was the easier side, in fact, it was not. The debate had no obvious winner.

5 Discussion

One goal of the workshop was to begin to concentrate the different threads of research in the IR community that touch on privacy into a coherent research agenda. To that end, the discussion was structured towards thinking of paper and session titles that we would like to see at SIGIR next year. (Hopefully, we like them enough to write the papers that will be presented in those sessions!)

Among those mentioned were:

- Helping users understand privacy policies.
- Private session/dynamic search.
- Detecting malicious links.
- Blending web and personal information search.
- Data markets for personal information.
- Detecting personal/sensitive information.
- Automatic redaction for privileged and classified information.
- Data challenge: get the most utility from this DP query log.

One proposal that received significant discussion was to identify a high-impact query log study, and define a differential privacy model for that study. For maximum effect, the paper could include the authors of the original study, and a released, privacy-preserving log. But the true impact of such work would be to show what other aspects of the log can be studied, and if there are questions answered by the original log that the DP log cannot support.

Another proposal from the workshop organizers was on transparency in personalized search. All commercial search results are heavily personalized, but the user can't easily recognize that. We propose augmenting the search results page with "personalization explanations" in an unobtrusive way. For example, imagine hovering the mouse over a search result link and seeing a pop-up explaining that you browsed that link two weeks ago. Another link might say that it is recommended because other searchers with similar queries clicked it. This would require a targeted, limited vocabulary of explanations that can be mapped to what are really complex ranking phenomena, but in a way that is helpful to the user. Finally, as a last step, imagine that the user, after seeing the explanation and realizing an unwanted leak of personal information, could click an X to indicate that the search engine should forget that information about the user.

Li Xiong proposed markets for personal and sensitive information. Your private information is out there, being bought and sold and used to optimize ranking functions. Companies have a role in that market, but you, the user, do not, apart from choosing what you type into your keyboard. If users could participate in the market by pricing and maintaining license control of their personal information, that would create an economic and legal mechanism for control of that information. Perhaps the biggest challenge is how to make such a market efficient.