# Poster: Searching HIE with Differentiated Privacy Preservation

Yuzhe Tang [§] Ling Liu [†]
[§]*Department of EECS, Syracuse University, Syracuse, NY, USA*
[†]*College of Computing, Georgia Institute of Technology, Atlanta, GA, USA*

*Abstract*—**In emerging Health Information Exchange systems (or HIE), a search facility, such as record locator service, is critically important for data sharing across autonomous hospitals. An understudied problem for searching HIE is the privacy preservation – how to protect the patient's private visit-history data in the search process and how to address innately different privacy and sensitivity for different patients and hospitals. For instance, knowing that a patient visited a specialty hospital (e.g. a women's health center) may leak more privacy than knowing that the patient visited a general hospital. In this work we proposed a differentiated privacy preservation technique for searching in HIE, coined PPLS. Given hospitals with different specialties, PPLS attempts to cluster them in order to hide among other hospitals their specialties linked to a patient, so that an attacker can not infer the patient's medical condition based on the specialties of the hospitals she visited.**

## I. Background

In the era of big data, human activities are computerized and are supported in the digital world. In the domain of Health IT, Health Information Exchange systems or HIE recently emerge (e.g. NHIN [1], GaHIN [2] and CommonWell [3]), in which patients delegate their digital medical records to the hospitals where they visited and the hospitals form a nation-wide (or state-wide) network to share information. Different hospitals are run autonomously and may compete for the same customer base (i.e. patients) which renders it difficult to establish full trust relationships between them. As regulated by Federal laws (e.g. HIPAA [4]), the hospitals are responsible for protecting patient privacy.

Information sharing is crucial for various applications in HIE. For example, when a patient who is unconscious is sent to a hospital, information sharing between multiple hospitals can help the doctor retrieve the patient's visit-history for immediate and accurate medical treatment. To establish information-sharing sessions, the Record Locator Service is a standard procedure defined by many HIE protocols [5], [1], [6], [7]; it provides the ability to discover where a patient's records are located based upon her identity, and is used as the first step towards establishing data-sharing relationship between the searcher and the patient's hospital of interest.

## II. Problem

It is desirable to protect privacy of a locator service in HIE. On the one hand, a locator service should maintain meta-data regarding the patients' visit history (i.e. the mapping from a patient to a hospital). Such meta-data is private and sensitive by itself; for example, the fact that a sports celebrity, say Mr. Tiger Woods, visited a hospital is something that he wants to keep confidential since disclosing it may jeopardize his future career. On the other hand, due to various practical concerns the locator service needs to be open[1], that is, to make the locator service accessible to all possible searchers, such as a licensed doctor who is a semi-honest human being at the same time. In this work, we consider a system in which the record locator service is hosted by a central third-party entity,[2] which is assumed to be untrusted. This assumption is made based on the fact that HIE involves autonomous hospitals which are mutually untrusted and it is difficult to find a central entity for all the hospitals to trust unanimously.[3]

This research addresses the privacy preservation of sensitive patient visit-history in the untrusted[4] locator service in HIE. Our unique observation is that privacy preservation should be *differentiated* when it comes to different patients and hospitals. Formally, the private visit-history data in HIE can be formulated to be that "a patient $t_j$ has visited (or has her records stored on) hospital $p_i$". Disclosing different private data raises privacy concerns of different levels. In particular, there are two kinds of differentiated privacy: 1) hospital-differentiated sensitivity and 2) personalized privacy. The hospital-differentiated sensitivity addresses the innate difference of sensitivity for different hospitals. For example, a woman may consider her visit to a women's health center (e.g., for an abortion) much more sensitive than her visit to a general hospital (e.g., for cough treatment). The personalized privacy addresses different patient's privacy concerns. For example, while an average person may not care too much about his/her visit to a hospital, a celebrity may be much more concerned about it, because even a small private matter of a celebrity can be publicized by the media (e.g., by paparazzi). It is therefore critical to differentiate privacy protection for different patients and hospitals. While our previous work $\epsilon$-PPI [9], [10] addresses the personalized privacy preservation, we in this work primarily focus on preserving the differentiated privacy by different hospitals.

A naive way to protect hospital-differentiated privacy is to anonymize the identity of a hospital by using $k$-anonymity; Specifically, hospitals are clustered into disjoint privacy groups of size $k$ in which a specific hospital is not distinguishable from any other hospitals in the same group. However, existing privacy-aware clustering approaches [11], [12] randomly assign hospitals to groups of uniformly equal size $k$, an approach that is agnostic to the different sensitivity of different hospitals.

## III. Proposed Approach

In this paper, we propose a Privacy-Preserving Locator Service, coined PPLS, to address the hospital-differentiated sensitivity. The key insight is that individual hospitals are

[1] The practical motivations include the needs of promoting data sharing and the difficulty of enforcing access controls without trusts

[2] This locator service may be distributed across multiple hospitals.

[3] One may argue that the U.S. government can serve as a candidate for the trusted entity. However, various scandals including the recent PRISM program [8] made the government lose the public trust.

[4] Currently, we assume locator service is untrusted only in terms of information confidentiality and user privacy, but is trusted in service and data integrity.

at different granularity in terms of their medical specialties. For example, Grady, a general hospital in Atlanta, has a variety of medical departments and thus knowing a patient visited Grady, a third party can not infer too much sensitive information regarding the patient's medical conditions. By contrast, Summit, a women's health center, is much more specific in its treatment specialty; knowing a woman patient visited Summit, the attacker may be much more confident about the woman's medical conditions (e.g. for an abortion). Thus, our PPLS technique never discloses information with certainty about a patient having visited a specialty hospital (e.g. Summit). To achieve this, as will be elaborated, our approach is to hide a hospital among other hospitals with different specialties.

### A. System Model

Our system model considers $n$ hospitals in $m$ medical specialty categories. Each hospital $h_i$ is modeled as a vector of $m$ elements, each being a rating from 0 to $k$ that represents the ranking of that hospital in the specialty category. In particular, if the rating is 0, it means that the hospital does not provide the specialty. In addition, each hospital maintains a list of patients who visited the hospital. We assume that a hospital's specialty areas are public (background) information. The patient list contains identifiable patient demographic information which is defined to be Protected Health Information (e.g. by HIPAA) and should be kept private.

In addition to $n$ hospitals, our system also contains a locator service presumably hosted by a third-party entity. The locator service bridges a searcher (e.g. a doctor) with hospitals that a patient of interest visited before. This search process, which facilitates the data sharing across hospitals, is modeled as a two-phase procedure. Specifically, as illustrated in Figure 1, a doctor who is interested in a certain patient first queries the locator service, from which the doctor obtains a list of hospitals that may be visited by the patient. Then for each hospital in the list, the doctor contacts it, locally searches and retrieves the private records of the patient after being authenticated and authorized.

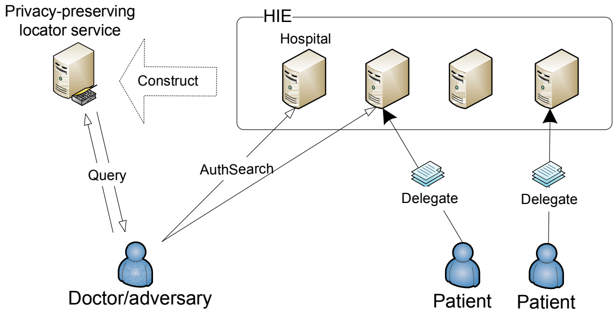

Fig. 1: HIE and privacy-preserving locator service

### B. Goal

Our goal is to protect the privacy of patient of interest in the search process. Our approach is based on the clustering technique in which hospitals clustered in a group are not distinguishable (from an outsider's point of view). After clustering, each group would have a merged vector of specialties, whose cardinality is called the group-wise specialty diversity (in a similar spirit to $l$-diversity [13] but adapted to the HIE context) which is currently considered as the most important factor for protecting patient privacy. Specifically, to a legitimate searcher, the clustering mechanism leads to extra search overhead: With locator service built at granularity of groups, the searcher needs to contact all hospitals in a group or none; thus s/he may be falsely redirected to hospitals that do not have the records of interest. Thus, it is desirable (from a legitimate searcher's perspective) to minimize the extra search overhead. On the other hand, to an adversary searcher, the clustering mechanism preserves privacy: By having hospitals with a variety of specialties in one group, a searcher for a patient can not infer sensitive information after he obtains the result from locator service and probably knows that the patient visited certain hospitals in a group. For instance, if all hospitals in the same group are with one and only one specialty, say Cancer treatment, then the searcher can easily infer that the patient is very likely to have cancer. But if the hospitals are with different specialties, then the searcher can not infer which diseases the patient has regarding his/her hospital visit.

### C. Construction Approach

To construct the PPLS, our proposed approach is a two-stage process. In the first stage, the hospitals are clustered based on their specialty vectors. Because a hospital's specialty is publicly known, this stage can be executed by an untrusted party where such specialty information is safe to release. We propose a greedy algorithm to efficiently cluster specialty vectors (of hospitals) to groups with diversity; the heuristic of the algorithm is to iteratively merge two existing vectors so that a maximal increase in specialty diversity can be obtained by merging them. Given a diversity goal, say $l < m$, the stop condition of the algorithm is that all groups are with specialties no less than $l$. In the second stage, each group employs a secure distributed computation protocol [14] to merge the sensitive patient lists from all member hospitals. Eventually, each group is related to a wide variety of specialties and a merged list of patients, both of which preserve privacy and are safe to release to the third-party locator service.

## ACKNOWLEDGEMENT

This work was partially supported by the National Science Foundation under Grants IIS-0905493, CNS-1115375, IIP-1230740 and Intel ISTC on Cloud Computing.

## REFERENCES

[1] "Nhin: http://www.hhs.gov/healthit/healthnetwork."
[2] "Gahin: http://www.gahin.org/."
[3] "Commonwell: http://www.commonwellalliance.org/."
[4] "Hipaa, http://www.cms.hhs.gov/hipaageninfo/."
[5] "Commonwell rls: http://www.commonwellalliance.org/services."
[6] "Nhin connect, http://www.connectopensource.org/."
[7] "Openempi: https://openempi.kenai.com/."
[8] "Prism, http://en.wikipedia.org/wiki/prism_(surveillance_program)."
[9] Y. Tang, L. Liu, A. Iyengar, K. Lee, and Q. Zhang, "e-ppi: Locator service in information networks with personalized privacy preservation," in *IEEE ICDCS 2014, Madrid, Spain, June 30 - July 3, 2014*, 2014, pp. 186–197.
[10] Y. Tang and L. Liu, "Privacy-preserving multi-keyword search in information networks," *TKDE 2015*.
[11] M. Bawa, R. J. B. Jr., and R. Agrawal, "Privacy-preserving indexing of documents on the network," in *VLDB*, 2003, pp. 922–933.
[12] Y. Tang, T. Wang, and L. Liu, "Privacy preserving indexing for ehealth information networks," in *CIKM*, 2011, pp. 905–914.
[13] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity," in *ICDE*, 2006, p. 24.
[14] L. Kissner and D. Song, "Privacy-preserving set operations," in *in CRYPTO 2005, LNCS*, 2005, pp. 241–257.