Poster: How well do you blend into the crowd?
-
d-convergence: A novel paradigm for quantifying privacy in the age of Big-Data

Michael Backes
CISPA, Saarland University
backes@cs.uni-saarland.de

Pascal Berrang
CISPA, Saarland University
berrang@cs.uni-saarland.de

Praveen Manoharan
CISPA, Saarland University
manoharan@cs.uni-saarland.de

I. Motivation

The Internet has undergone dramatic changes in the last two decades, evolving from a mere communication network to a global multimedia platform in which billions of users not only actively exchange information, but increasingly conduct sizable parts of their daily lives. While this transformation has brought tremendous benefits to society, it has also created new threats to online privacy that existing technology is failing to keep pace with. Users tend to reveal personal information without considering the widespread, easy accessibility, potential linkage and permanent nature of online data. Many cases reported in the press show the resulting risks, which range from public embarrassment and loss of prospective opportunities (e.g. when applying for jobs or insurance), to personal safety and property risks (e.g. when sexual offenders or burglars learn users' whereabouts online). The resulting privacy awareness and privacy concerns of Internet users have been further amplified by the advent of the Big-Data paradigm and the aligned business models of personalized tracking and monetizing personal information in an unprecedented manner.

Developing a suitable methodology to reason about Big-Data privacy, as well as corresponding tool support in the next step, requires at its core a formal privacy model for assessing and quantifying to what extent a user is disseminating private information on the Internet. Any adequate privacy model needs to live up to the now increasingly dynamic dissemination of unstructured, heterogeneous user content on the Internet: While users traditionally shared information mostly using public profiles with static information about themselves, nowadays they disseminate personal information in an unstructured, highly dynamic manner, through content they create and share (such as blog entries, user comments, a "Like" on Facebook), or through the people they befriend or follow. Furthermore, ubiquitously available background knowledge about a dedicated user needs to be appropriately reflected within the model and its reasoning tasks, as it makes it possible to decrease a user's privacy by inferring further sensitive information. As an example, Machine Learning and other Big-Data analysis techniques provide comprehensive approaches for profiling a user's actions across multiple online social networks, up to a unique identification of a given user's profiles for each such network.

II. Inadequacy of existing Models

As of now, *even the basic methodology is missing* for offering users technical means to comprehensively assess the privacy risks incurred by their data dissemination, and their daily online activities in general. Existing privacy models such as k-anonymity [1], l-diversity [2], t-closeness [3] and the currently most popular notion of Differential Privacy [4] follow a database-centric approach that is inherently inadequate to meet the requirements outlined in the previous paragraph: these notions require a) an a priori sensitivity assessment of personal information which can be highly context sensitive on the web, b) a pre-defined structure on data, whereas web data is very heterogeneous and unstructured and c) global sanitation of the whole data set, which is simply impossible in the online setting, where we can, at most, only sanitize the single user's input into the system.

III. Model for Online Privacy

We develop a novel formal privacy model that addresses the above mentioned shortcomings of existing approaches. It is based on the concept of statistical language models, which is the predominantly used technique in the Information Retrieval (IR) community for characterizing documents with regard to their information content [5], [6]. Grounding our model upon such statistical models allows us to cope with unstructured, heterogeneous data, as well as highly dynamic content generation. Moreover, it allows us to seamlessly incorporate future advances from IR research and other Big-Data technologies into our model.

Our model defines and quantifies privacy by utilizing the notion of entity similarity, i.e., an entity is private in a collection of entities if it is sufficiently similar to its peers. Formally, this intuition is captured by defining corresponding statistical models that allow us to characterize entities based

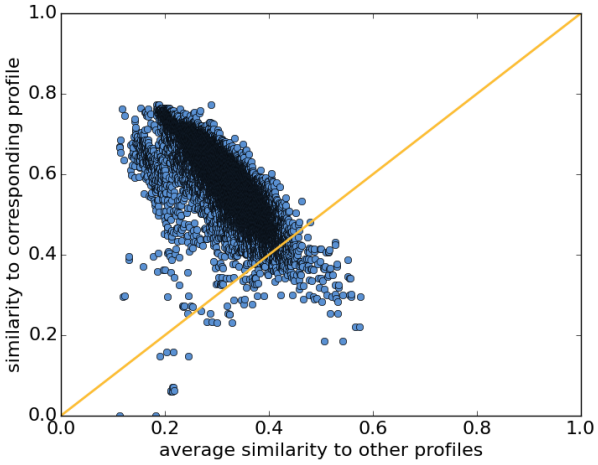

Figure 1: The average similarity between $P_{u,s}$ and all profiles in $\Pi_{u,s'}$ versus the matching value between $P_{u,s}$ and $P_{u,s'}$.

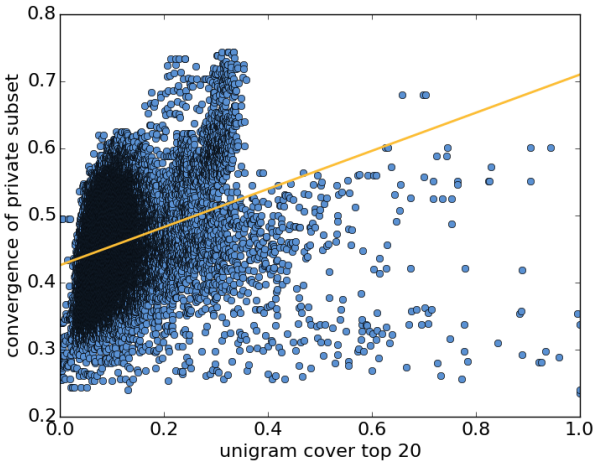

Figure 2: The top 20 unigram cover of a profile $P_{u,s}$ versus the convergence of its private subset $\Pi_{u,s}$.

on the information they have disseminated publicly and based on ubiquitously available background knowledge about these entities. At the technical core of our model is the new notion of d-convergence, which measures the similarity of entities within a larger group of entities. It hence provides the formal grounds to quantify the ability of any single entity to blend into the crowd, i.e., to hide amongst peers. In contrast to existing models, we do not have to differentiate between non-sensitive and sensitive attributes, but rather start from the assumption that all data is equally important and can lead to privacy risks. More specifically, our model captures the fact that the sensitivity of attributes is highly context-dependent, i.e., attributes can be or become sensitive for a specific entity when interacting with its peers. We furthermore show how to leverage our model for identifying context-specific, privacy-critical attributes.

We show that our model and its underlying notion of d-convergence implies existing privacy notions if one considers a setting with *structured* data only: we define a suitable transformation of our statistical model to a statistical database and subsequently show that a d-convergent database is also t-close.

Our privacy model is furthermore capable of assessing privacy risks specifically for single entities. To this end, we extend the notion of d-convergence to the novel notion of (k, d)-privacy, which allows for entity-centric privacy assessments by requiring d-convergence in the local neighborhood of a given entity. This definition thus allows us to make user-centric privacy assessments and provide lower bounds for an individual user's privacy irrespective of the whole data set, i.e., these bounds stay valid even when the set is enlarged, e.g., by including new users.

IV. Experimental Evaluation

We present an instantiation of our general privacy model for the important use case of analyzing user-generated text content in order to characterize specific user profiles. We use unigram frequencies extracted from user-generated content as user attributes, and we subsequently demonstrate that the resulting unigram model can indeed be used for quantifying the degree of anonymity of—and ultimately, for differentiating—individual entities.

We apply this unigram model to a collection of 40 million comments collected from the Online Social Network Reddit, which we stripped down to 15 million comments to keep the evaluation tractable. The computations were performed on two Dell PowerEdge R820 with 64 virtual cores each at 2.60GHz over the course of six weeks. Our evaluation show that the statistical model approach is suited for modeling users (see Figure 1) and confirm hypotheses about the identifiability of entities in our dataset, e.g. entities that conform to the collection's average behavior are more difficult to identify (see Figure 2). We thereby validate our statistical model approach for evaluating privacy characteristics in real world settings.

References

[1] L. Sweeney, "K-Anonymity: A Model for Protecting Privacy," *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, vol. 10, no. 5, pp. 557–570, 2002.

[2] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "L-Diversity: Privacy Beyond K-anonymity," *ACM Trans. Knowl. Discov. Data*, vol. 1, no. 1, 2007.

[3] N. Li and T. Li, "t-closeness: Privacy beyond k-anonymity and -diversity," in *In Proceedings of the 23rd International Conference on Data Engineering (ICDE*, 2007.

[4] C. Dwork, "Differential Privacy: A Survey of Results," in *Proceedings of the 5th International Conference on Theory and Applications of Models of Computation*, 2008, pp. 1–19.

[5] J. M. Ponte and W. B. Croft, "A Language Modeling Approach to Information Retrieval," in *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1998, pp. 275–281.

[6] V. Lavrenko and W. B. Croft, "Relevance based language models," in *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2001, pp. 120–127.