Poster: How well do you blend into the crowd?

\textit{d}-convergence: \mbox{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.

II. Inadequacy of Existing Models

As of now, \textit{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 \textit{k} anonymity \cite{1}, \textit{l} diversity \cite{2}, \textit{t} closeness \cite{3} and the currently most popular notion of Differential Privacy \cite{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 \cite{5}, \cite{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

\textit{d}-convergence: \mbox{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.

II. Inadequacy of Existing Models

As of now, \textit{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 \textit{k} anonymity \cite{1}, \textit{l} diversity \cite{2}, \textit{t} closeness \cite{3} and the currently most popular notion of Differential Privacy \cite{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 \cite{5}, \cite{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
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