Skysong Innovations
Advancing the Arizona State University Knowledge Enterprise

1475 N. Scottsdale Road, Suite 200
Scottsdale, AZ 85287-3538
Phone: 480 884 1996 Fax: 480 884 1984

Case ID:M18-252P^
Published: 2/26/2020

Inventors

Yanchao Zhang

Jinxue Zhang

Contact

Shen Yan
shen.yan@skysonginnovations.com

Privacy-Preserving Social Media Data Outsourcing

Background

User-generated social media data is growing exponentially along with the demand for its use in identifying user trends and behavior. As data is outsourced from social networks to data consumers, there exists a fundamental tradeoff between data utility and user privacy. Transferring intact user data is most profitable for the provider and valuable for the consumer, but risks exposure of sensitive user information. This includes inferred information that is not explicitly disclosed, such as age, location, and political preferences.

Anonymizing user identities prior to outsourcing has been one way to increase privacy. However, malicious data consumers may still be able to link random or anonymous identities within an outsourced data set to real identities on a social media platform. Often enabling these user-linkage attacks are the user text data (e.g., the content of a tweet) that remain unobfuscated. Because user text data is considered critical to the value of an outsourced data set, preventing these attacks while preserving data utility remains a key challenge.

Invention Description

Researchers at Arizona State University have developed a new framework whereby user-generated text is sufficiently obfuscated and disclosed along with anonymized identities. This method involves the mapping of intact data of all users to a high-dimensional user-keyword matrix, to which controlled noise is added to achieve differential privacy. A novel text-based characterization process is introduced that effectively bypasses the so-called "Curse of Dimensionality," a limitation encountered by popular Laplacian approaches to differential privacy.

The testing of a real-world Twitter data set demonstrated high-level privacy protection with minimal sacrifice of utility. Privacy leakage was reduced by as much as 64.1% with only a 1.61% decrease in classification accuracy.

Potential Applications

- Social media networks

- Third-party data providers
- Demographic reporting

Benefits and Advantages

- Utility-preserving – Minimal effect on data mining task results
- Privacy-preserving – Text-based user-linkage attacks are rendered virtually impossible
- Data-preserving – All user data maps to matrix without additions or deletions