Article By Cristian Augusto et al., Artículo presentado en las XXIV Jornadas de Ingeniería del Software y Bases de Datos (JISBD 2019) celebradas en Cáceres del 2 al 4 de septiembre de 2019, como parte de las Jornadas SISTEDES. Se permite cualquier explotación de la obra, incluyendo una finalidad comercial, así como la creación de obras derivadas, la distribución de las cuales también está permitida sin ninguna restricción.

# Anonimización de datos guiada por pruebas para aplicaciones inteligentes

Cristian Augusto, Jesús Morán, Claudio de la Riva, Javier Tuya

Departamento de Informática, Universidad de Oviedo, Gijón,
{augustocristian, moranjesus, claudio, tuya}@uniovi.es

**Resumen.** En la actualidad gran cantidad de datos son compartidos para su uso, tratamiento o análisis entre empresas y terceros. Es habitual que estos datos tengan que ser protegidos con diferentes técnicas de preservación de la privacidad para dar cumplimiento a las leyes y regulaciones. Una de las técnicas más comunes es la anonimización que, aunque provee de privacidad a los datos, presenta como efecto colateral la perdida de información. Esta pérdida de información puede afectar negativamente al comportamiento de aquellos desarrollos altamente dependientes de dichos datos como son las aplicaciones inteligentes. Para abordar este problema, proponemos un enfoque guiado por pruebas para seleccionar el conjunto de datos anonimizado que mantenga un compromiso entre la calidad no funcional (privacidad) y la funcional (utilidad). Para ello se alimenta a las aplicaciones con los datos anonimizados para que tomen los patrones de comportamiento de estos, y seguidamente validar las predicciones con los datos originales, midiendo así su calidad funcional. Dicha calidad junto con la no funcional (privacidad), es ponderada según los criterios de usuario con el fin de alcanzar el punto de compromiso entre ambas características de calidad.

**Palabras clave:** Anonimización, Pruebas de Software, Aplicaciones Inteligentes, Privacidad

## 1 Introducción

En muchas ocasiones las empresas transfieren grandes cantidades de sus datos a terceros para los puedan consultar y crear aplicaciones (como por ejemplo utilidades para diagnosticar de forma automática enfermedades o deducir el índice de riesgo de impago de un crédito hipotecario). Varios de estos datos pueden contener información sensible que se debe proteger debido a aspectos legales externos o internos, como pueden ser datos médicos de un paciente. Los proveedores que externalicen/distribuyan/comparten sus datos con terceros o los liberen siguiendo la filosofía del "Open Data" deberán proteger estos datos con técnicas como la anonimización para evitar la asociación de datos sensibles con un individuo[1]. Con la anonimización de datos se suprime o se altera información para evitar la asociación de datos sensibles con un individuo, por lo que se puede perder información que afecte a la funcionalidad de las aplicaciones que empleen estos datos alterados, como ocurre con las aplicaciones inteligentes.

Article By Cristian Augusto et al., Artículo presentado en las XXIV Jornadas de Ingeniería del Software y Bases de Datos (JISBD 2019) celebradas en Cáceres del 2 al 4 de septiembre de 2019, como parte de las Jornadas SISTEDES. Se permite cualquier explotación de la obra, incluyendo una finalidad comercial, así como la creación de obras derivadas, la distribución de las cuales también está permitida sin ninguna restricción.



Si bien los proveedores de datos deben asegurar la privacidad de la información (calidad no funcional) mediante la anonimización, también deben proporcionar datos útiles que no afecten negativamente a la funcionalidad de las aplicaciones que los utilizan (calidad funcional). Este aspecto cobra especial importancia en el caso de las aplicaciones inteligentes ya que si los datos de los que aprenden son incorrectos, debido a que estén demasiado alterados por la anonimización, los pronósticos y predicciones también lo estarán. Es por ello por lo que el proveedor antes de liberar los datos debería realizar pruebas que garanticen que los datos anonimizados por un lado preserven la privacidad (calidad no funcional) y que sigan siendo útiles para las aplicaciones inteligentes (calidad funcional).

El presente artículo propone un enfoque guiado por pruebas para que la anonimización consiga un compromiso entre ambas calidades, es decir que los datos preserven la privacidad y que además sean útiles para el desarrollo de las aplicaciones inteligentes. Para ello se realizan pruebas de varias anonimizaciones, en busca de aquella que maximice ambas calidades teniendo en cuenta las preferencias del usuario (darle más prioridad a la utilidad o a la privacidad de los datos anonimizados).

Diversos autores han estudiado la posibilidad de utilizar datos anonimizados para desarrollar aplicaciones inteligentes [2, 3]. La mayor parte de trabajos se centra en maximizar únicamente uno de los atributos de calidad (calidad no funcional y funcional, respectivamente) [4, 5]. En cambio, nuestro enfoque plantea un compromiso entre calidad funcional (calidad de los datos) y la no funcional (privacidad) teniendo en cuenta las preferencias del usuario, empleando los datos reales para probar el comportamiento de dichas aplicaciones inteligentes.

Las contribuciones de este articulo son:

1. Un enfoque de anonimización de datos guiada por pruebas del software para lograr un compromiso entre la privacidad de los datos y la calidad funcional de las aplicaciones inteligentes que aprenden de dichos datos.
2. Una métrica sobre la calidad de la anonimización compuesta de ambas calidades (funcional y no funcional) priorizadas por las preferencias de usuario.

## 2 Anonimización basada en pruebas

El enfoque propuesto consiste en realizar diferentes pruebas sobre las operaciones de anonimización (en adelante referida por anonimizaciones) para que los datos, además de preservar la privacidad, sean útiles para las aplicaciones. Por ello, en este caso SUT (*System Under Test*) será el conjunto de datos que tenemos que anonimizar. La problemática a abordar se ilustra con el siguiente ejemplo:

***Ejemplo:*** En Fig. 1 se muestra un conjunto de datos sin anonimizar (azul) y tres diferentes anonimizaciones de los mismos (en color rojo), que se emplean para predecir el valor de la función Y = f (X). En la primera grafica (A), el nivel de privacidad es bajo (1-Anonimidad) ya que se realiza sobre los datos originales, y presenta una buena calidad funcional ($R^2 = 0.979$). En el segundo modelo (B), se mejora ligeramente la privacidad (5-Anonimidad), pero su calidad funcional respecto al original empeora ligeramente ($R^2 = 0.953$). Tanto en el tercer como en el cuarto modelo (C y D) se vuelve

Article By Cristian Augusto et al., Artículo presentado en las XXIV Jornadas de Ingeniería del Software y Bases de Datos (JISBD 2019) celebradas en Cáceres del 2 al 4 de septiembre de 2019, como parte de las Jornadas SISTEDES. Se permite cualquier explotación de la obra, incluyendo una finalidad comercial, así como la creación de obras derivadas, la distribución de las cuales también está permitida sin ninguna restricción.



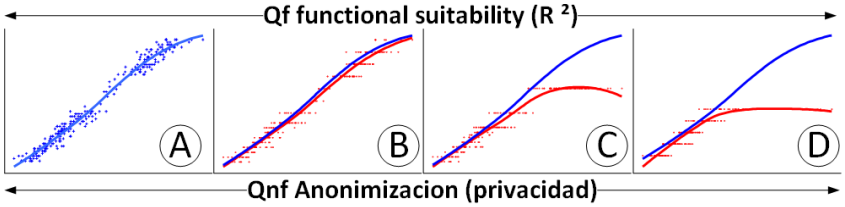


Fig. 1. Ajuste modelos anonimizado contra el modelo real

a mejorar sustancialmente la privacidad (10-Anonimidad y 40-Anonimidad, respectivamente). A medida que la privacidad aumenta, se empieza a observar el progresivo empeoramiento de la calidad funcional respecto a los datos originales ($R^2 = 0.11$). Sin embargo, respecto a los datos anonimizados se sigue manteniendo un buen ajuste ($R^2 = 0.92$), causando que las aplicaciones desarrolladas con estos den muy buenos resultados en desarrollo y acaben fallando dramáticamente al ser puestas en producción en un entorno real.

El enfoque propuesto, persigue maximizar la calidad del SUT ($Q$) ponderando su calidad funcional y no funcional (privacidad). La calidad funcional ($Q_F$) se evalúa a través de la calidad de predicción de la aplicación con los datos originales, utilizando una medida como puede ser $R^2$, precisión o la exactitud. Para la evaluación de la calidad no funcional ($Q_{NF}$) se emplean también medidas estándar como k-Anonimidad [6] o *l-diversity* [7] entre otras. El compromiso entre ambas calidades vendrá dado por una métrica de calidad $Q = Q_F + \alpha * Q_{NF.}$, siendo $\alpha$ un valor dado por las preferencias del usuario para priorizar una calidad frente a otra, de acuerdo con las necesidades de calidad funcional y privacidad del dataset. Nuestro enfoque realiza pruebas sobre diferentes anonimizaciones de los datos evaluando su calidad y quedándose con el dataset anonimizado con mayor calidad (Q) que dependerá de las preferencias del usuario.

En Fig. 2 se muestran los datasets anonimizados representados en Fig.1 (A, B, C y D) para diferentes preferencias del usuario: priorizando la calidad funcional (izquierda $\alpha = 0.125$), ponderando ambas calidades por igual (centro $\alpha = 0.25$) y finalmente dando más importancia a la calidad no funcional (derecha $\alpha = 0.5$). La anonimización que maximiza la calidad de las dos primeras gráficas (Fig.2 izquierda y centro) es B, presentando una calidad funcional alta ($R^2$=0.953), y una calidad no funcional de 5-Anonimidad. En la tercera gráfica la anonimización que maximiza la calidad es D, presentando un nivel de privacidad alto (40-Anonimidad) y una calidad funcional baja ($R^2$=0.11).

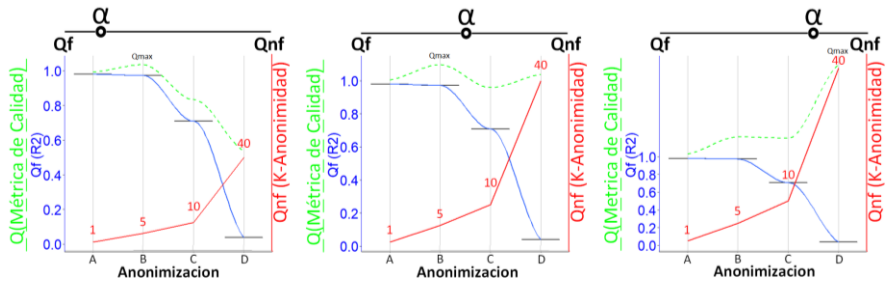


Fig. 2. Metrica de calidad dependiendo de los preferencias de usuario de los SUT Fig. 1

Article By Cristian Augusto et al., Artículo presentado en las XXIV Jornadas de Ingeniería del Software y Bases de Datos (JISBD 2019) celebradas en Cáceres del 2 al 4 de septiembre de 2019, como parte de las Jornadas SISTEDES. Se permite cualquier explotación de la obra, incluyendo una finalidad comercial, así como la creación de obras derivadas, la distribución de las cuales también está permitida sin ninguna restricción.



## 3 Conclusiones y Trabajo Futuro

Se ha introducido un enfoque de anonimización guiado por pruebas destinado a la transferencia de datos útiles para el desarrollo de aplicaciones inteligentes, preservando la privacidad de los datos sensibles. Con la transferencia útil y segura de los datos, estos pueden ser utilizados por la comunidad, permitiendo dar cabida a desarrollos de mucha mayor calidad además de fomentar iniciativas como el denominado *"open data"*.

Como trabajo futuro se propone comprobar la efectividad de dicho enfoque en más casos reales, tarea que ya se ha empezado a realizar [8], analizar la dependencia de la anonimización con las aplicaciones inteligentes que emplean dichos datos o automatizar el proceso de anonimización para maximizar el compromiso entre ambas calidades.

**Agradecimientos**
Este trabajo ha sido realizado bajo el proyecto de investigación TestEAMos (TIN2016-76956-C3-1-R), financiado por el Ministerio Español de Economía y Competitividad junto con fondos FEDER

**Referencias**

1. Fung, B.C.M., Wang, K., Chen, R., Yu, P.S.: Privacy-preserving Data Publishing: A Survey of Recent Developments. ACM Comput Surv. 42, 14:1–14:53 (2010). https://doi.org/10.1145/1749603.1749605.
2. Brickell, J., Shmatikov, V.: The cost of privacy: Destruction of data-mining utility in anonymized data publishing. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 70–78 (2008). https://doi.org/10.1145/1401890.1401904.
3. Buratović, I., Miličević, M., Žubrinić, K.: Effects of data anonymization on the data mining results. In: 2012 Proceedings of the 35th International Convention MIPRO. pp. 1619–1623 (2012).
4. Iyengar, V.S.: Transforming Data to Satisfy Privacy Constraints. In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 279–288. ACM, New York, NY, USA (2002). https://doi.org/10.1145/775047.775089.
5. Xu, J., Wang, W., Pei, J., Wang, X., Shi, B., Fu, A.W.-C.: Utility-based anonymization using local recoding. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 785–790. ACM (2006).
6. Samarati, P.: Protecting respondents identities in microdata release. IEEE Trans. Knowl. Data Eng. 13, 1010–1027 (2001). https://doi.org/10.1109/69.971193.
7. Machanavajjhala, A., Gehrke, J., Kifer, D., Venkitasubramaniam, M.: L-diversity: privacy beyond k-anonymity. In: 22nd International Conference on Data Engineering (ICDE'06). pp. 24–24 (2006). https://doi.org/10.1109/ICDE.2006.1.
8. Augusto, C., Morán, J., de la Riva, C., Tuya, J.: Test-driven Anonymization for Artificial Intelligence. In: 2019 IEEE International Conference on Artificial Intelligence Testing (AITest). pp. 103–110 (2018). https://doi.org/10.1109/AITest.2019.00011.