PROYECTO DE INVESTIGACIÓN INTERNOS SIN FINANCIAMIENTO O AUTOGESTIONADOS

ANEXO 1 - DATOS INFORMATIVOS

Fecha de presentación (dd/mm/aa): 07/08/2018

Título del proyecto:

Privacidad Sintáctica Funcional: Análisis y adaptación de mecanismos de anonimato con enfoque en la preservación de utilidad de los datos

TIPOS DE INVESTIGACIÓN

Investigación básica ❑ Investigación aplicada ■

DEPARTAMENTO(S) Y/O INSTITUTO(S):
1. Departamento de Electrónica, Telecomunicaciones y Redes de Información
2.

LÍNEA(S) DE INVESTIGACIÓN (verificable en el SAEW):
1. Seguridad y Privacidad
2.

RESUMEN DE INFORMACIÓN DEL DIRECTOR Y COLABORADORES

Director

Apellidos y nombres	No. de Cédula	HSS	Departamento	Título de mayor nivel y mención.
Urquiza Aguiar Luis Felipe	1717545287	6	Departamento de Electrónica, Telecomunicaciones y Redes de Información	Ph.D. en Ingeniería Telemática

Colaborador(es) Externos

Apellidos y nombres	No. de Cédula	HSS	Departamento	Título de mayor nivel y mención.
Rodríguez Hoyos Ana Fernanda	1713027157	N/A	Departamento de Electrónica, Telecomunicaciones y Redes de Información	MSc. en Ingeniería Telemática
Estrada Jiménez José Antonio	1714672456	N/A	Departamento de Electrónica, Telecomunicaciones y Redes de Información	MSc. en Ingeniería Telemática

* HSS = Horas Semana Semestre

HOJA DE VIDA DEL DIRECTOR DEL PROYECTO

Datos Personales					
Nombre completo:	Luis Felipe Urquiza Aguiar				
No. de identificación:	1717545287	**Nacionalidad:**	Ecuatoriano		
Fecha de nacimiento:	03/05/1983	**Celular:**	0996615630	**Ext. EPN:**	2311
Correo institucional:	luis.urquiza@epn.edu.ec				
Cargo actual en la EPN:	Profesor Titular Auxiliar				
Facultad:	Eléctrica y Electrónica				
Departamento:	Telecomunicaciones y Redes de la Información				

Educación universitaria. Proveer el nombre de los títulos de pregrado y postgrado (Ing., M.Sc., Ph.D.)				
Título	**Año**	**Institución/Universidad**	**Ciudad/País**	**Área o línea de investigación de la tesis**
Ingeniero en Electrónica y Redes Información	2008	Escuela Politécnica Nacional	Quito/Ecuador	Redes de área extendida, Calidad de servicio (QoS)
Máster en Ingeniería Telemática	2012	Universitat Politécnica de Catalunya	Barcelona/España	Redes ad hoc vehiculares, privacidad de información
Doctor en Ingeniería Telemática	2016	Universitat Politécnica de Catalunya	Barcelona/España	Redes ad hoc vehiculares, protocolos de encaminamiento
Máster en Estadística e Investigación Operativa	2018	Universitat Politécnica de Catalunya	Barcelona/España	Algoritmo de punto interior para ubicación óptima de videos.

Experiencia investigativa y en ejecución de proyectos (cite los tres más relevantes)		
Año	**Título del proyecto**	**Posición /Actividades realizadas**
01/01/2014 - 01/05/2016	"E-iRoads: Ecuador - Inteligent Roads. Un Sistema inteligente para la gestión de tráfico en las periferias de grandes ciudades (Caso de Estudio: Quito)	Investigador colaborador / Revisión de artículos científicos generados por otros miembros del grupo. Desarrollo de trazas de movilidad realistas para las vías de acceso a Quito.
01/01/2014 - 01/05/2016	Modelamiento de la Packet Error Rate (PER) incluyendo condiciones de Peak-to-Average Power Ratio (PAPR) para transmisiones Ad-Hoc.	Investigador colaborador/ Revisión de artículos científicos generados por otros miembros del grupo. Análisis y modificación del código del simulador NS-3 para incluir condiciones de PAPR
01/01/2014 - 01/05/2016	INcident MitoRing In Smart COmunities	Investigador / Revisión de artículos científicos generados por otros miembros del grupo. Desarrollo de protocolos de encaminamiento para comunicaciones anycast en VANETs
01/01/2014 - 30/06/2015	Ayuda puente: EMergency Response In Smart COmunities (EMRISCO)	Investigador / Estudio de Calidad de servicio en comunicaciones para una ciudad inteligente

Publicaciones, patentes, prototipos o productos (cite las más relevantes dentro de los últimos cinco años y que se encuentren alineados al proyecto de investigación)	
1.	Rodríguez-Hoyos, A., Estrada-Jiménez, J., Urquiza-Aguiar, L., Parra-Arnau, J., & Forné, J. (2018, April). Digital Hyper-Transparency: Leading e-Government Against Privacy. In eDemocracy & eGovernment (ICEDEG), 2018 International Conference on (pp. 263-268). IEEE

2.	Tripp-Barba, C., Urquiza-Aguiar, L., Aguilar-Igartua, M., Parra -Arnau, J., Rebollo-Monedero, D., Forné, J., & Pallarès, E. (2013). A collaborative protocol for anonymous reporting in vehicular ad hoc networks. Computer Standards & Interfaces, (0). https://doi.org/http://dx.doi.org/10.1016/j.csi.2013.06.001
3	Rebollo-Monedero, D., Forné, J., Pallarès, E., Parra -Arnau, J., Tripp, C., Urquiza, L., & Aguilar, M. (2014). On collaborative anonymous communications in lossy networks. Security and Communication Networks, 7(12), 2761–2777. https://doi.org/10.1002/sec.793
4.	Urquiza, L., Rodríguez, A., Tripp, C., & Aguilar, M. (2016). Evaluation of VANET Performance for Anonymous Reporting through Coherent , Automatic Address Resolution (CAAR). Revista Politécnica, 37(1), 1–8. Retrieved from http://www.revistapolitecnica.epn.edu.ec/ojs2/index.php/revista_politecnica2/article/view/567

Experiencia profesional, otros trabajos científicos y técnicos (cite lo más relevante o las más recientes)

Función/Cargo: Profesor Titular Auxiliar
Institución: Escuela Politécnica Nacional
País / Ciudad: Ecuador / Quito
Período: 1 de septiembre de 2017 -actual
Actividades Preparación, dictado, revisión de exámenes de materias, participación proyectos de investigación

Función/Cargo: Profesor Ocasional
Institución: Escuela Politécnica Nacional
País / Ciudad: Ecuador / Quito
Período: 1 de julio de 2016 – 31 de agosto de 2017
Actividades Preparación, dictado, revisión de exámenes de materias, participación proyectos de investigación

Función/Cargo: Asistente de cátedra
Institución: Escuela Politécnica Nacional
País / Ciudad: Ecuador / Quito
Período: 1 de marzo a 30 de agosto de 2010
Actividades Preparación, dictado, revisión de exámenes de materias

Función/Cargo: Asistente de Investigación
Institución: Departamento de Ingeniería Telemática – Universitat Politècnica de Catalunya
País / Ciudad: España / Barcelona
Período: 1Febrero 2011 – 5 de septiembre 2012 y 1 de enero de 2014 – 1 de mayo de 2016
Actividades Participación en las actividades del grupo de investigación SeRvicios TELemáticos.

Función/Cargo: Asistente de cátedra
Institución: Escuela Politécnica Nacional
País / Ciudad: Ecuador / Quito
Período: 1 de marzo a 30 de agosto de 2010
Actividades Preparación, dictado, revisión de exámenes de materias

Función/Cargo: Ingeniero Servicios Tecnológicos
Institución: Unidad de Ejecución Especializada/ Ministerio de Gobierno del Ecuador
País / Ciudad: Ecuador / Quito
Período: 1 de julio de 2008 a 5 de marzo de 2010
Actividades: Planificación de actividades de soporte a proyectos e instalaciones.

Función/Cargo: Técnico
Institución: Cuerpo de Ingenieros del Ejército / Petrocomercial
País / Ciudad: Ecuador / Quito
Período: 1 de diciembre 2007 a 1 de julio 2008
Actividades: Participación en el diseño e implementación de proyectos de tecnología.

ESCUELA POLITÉCNICA NACIONAL
VICERRECTORADO DE INVESTIGACIÓN Y PROYECCIÓN SOCIAL

HOJA DE VIDA DEL PROFESOR COLABORADOR EXTERNO DEL PROYECTO (1)

Datos Personales						
Nombre Completo:	Ana Fernanda Rodríguez Hoyos					
No. de Identificación:	1713027157	**Nacionalidad:**	Ecuatoriana			
Fecha de nacimiento:	02/12/1985	**Celular:**	+34630950193	**Ext. EPN:**		
Correo institucional:	ana.rodriguez@epn.edu.ec					
Cargo Actual en la EPN:	Profesor Agregado a Tiempo Completo					
Facultad:	Facultad de Ingeniería Eléctrica y Electrónica					
Departamento:	Departamento de Electrónica, Telecomunicaciones y Redes de Información					

Educación universitaria. Proveer el nombre de los títulos de pregrado y postgrado (Ing., Magister, Ph.D.)				
Título	**Año**	**Institución/Universidad**	**Ciudad/País**	**Área o línea de investigación de la tesis**
Magister	2013	Universidad Politécnica de Cataluña	Barcelona/España	Seguridad y Privacidad
Ing.	2010	Escuela Politécnica Nacional	Quito/Ecuador	Servicios de Red

Experiencia investigativa y en ejecución de proyectos (cite los tres más relevantes)		
Año	**Título del proyecto**	**Cargo /Actividades realizadas**
2014-2016	Desarrollo de Prototipos para la Investigación en Seguridad de la Información	Investigador Colaborador

Publicaciones, patentes, prototipos o productos (cite las más relevantes dentro de los últimos cinco años y que se encuentren alineados al proyecto de investigación)	
1.	Rodríguez-Hoyos, A., Estrada-Jiménez, J., Rebollo-Monedero, D., Parra-Arnau, J., & Forné, J. (2018). Does k-Anonymous Microaggregation Affect Machine-Learned Macrotrends?. IEEE Access.
2.	Rodríguez-Hoyos, A., Estrada-Jiménez, J., Urquiza-Aguiar, L., Parra-Arnau, J., & Forné, J. (2018, April). Digital Hyper-Transparency: Leading e-Government Against Privacy. In eDemocracy & eGovernment (ICEDEG), 2018 International Conference on (pp. 263-268). IEEE
3.	Estrada-Jiménez, J., Parra-Arnau, J., Rodríguez-Hoyos, A., & Forné, J. (2017). Online advertising: Analysis of privacy threats and protection approaches. Computer Communications, 100, 32-51.
4.	Estrada, J. A., Estrada, J. C., Rodríguez, A. F., & Tipantuña, C. J. (2015). Ecuador y la Privacidad en Internet: Una Aproximación Inicial. Revista Politécnica, 36(1), 54.
5.	Estrada-Jiménez, J. A., Rodríguez, A. F., Parra, J., & Forné, J. (2014). Evaluation of a Query-Obfuscation Mechanism for the Privacy Protection of User Profiles. Network Protocols and Algorithms, 6(2), 55-92.

Experiencia profesional, otros trabajos científicos y técnicos (cite lo más relevante o las más recientes)
• Escuela Politécnica Nacional – EPN Profesora a tiempo completo de las asignaturas: Sistemas Microprocesados, Probabilidad y Estadística, Programación y Software de Simulación. Investigadora en áreas relacionadas con Seguridad y Privacidad de la Información Agosto 2013 – actualidad • Universidad de las Américas (UDLA) Profesora de la asignatura Tecnologías de Última Milla Marzo 2014 – Abril 2014 • Ministerio del Interior Administradora de Infraestructura Tecnológica (Servidor Público 3). Configuración y mantenimiento de servidores de correo electrónico, proxy-firewall, e infraestructura inalámbrica.

Abril 2011 – Agosto 2011

- Escuela Politécnica Nacional – EPN – DETRI
  Ayudante de Laboratorio del Departamento de Electrónica Telecomunicaciones y Redes de Información de la Escuela Politécnica Nacional. Mantenimiento, administración y soporte de los laboratorios de redes y comunicación digital. Instructora del laboratorio de Comunicación Digital y Dispositivos Electrónicos de la carrera de Electrónica y Telecomunicaciones
  Septiembre 2009 - Febrero 2011.

HOJA DE VIDA DEL PROFESOR COLABORADOR EXTERNO DEL PROYECTO (2)

Datos Personales						
Nombre Completo:	José Antonio Estrada Jiménez					
No. de Identificación:	1714672456	**Nacionalidad:**	Ecuatoriana			
Fecha de nacimiento:	29/06/1983	**Celular:**	+34657686858	**Ext. EPN:**		
Correo institucional:	jose.estrada@epn.edu.ec					
Cargo Actual en la EPN:	Profesor Agregado a Tiempo Completo					
Facultad:	Facultad de Ingeniería Eléctrica y Electrónica					
Departamento:	Departamento de Electrónica, Telecomunicaciones y Redes de Información					

Educación universitaria. Proveer el nombre de los títulos de pregrado y postgrado (Ing., Magister, Ph.D.)				
Título	**Año**	**Institución/Universidad**	**Ciudad/País**	**Área o línea de investigación de la tesis**
Magister	2013	Universidad Politécnica de Cataluña	Barcelona/España	Seguridad y Privacidad
Ing.	2007	Escuela Politécnica Nacional	Quito/Ecuador	Servicios de Red

Experiencia investigativa y en ejecución de proyectos (cite los tres más relevantes)		
Año	**Título del proyecto**	**Cargo /Actividades realizadas**
2014-2016	Desarrollo de Prototipos para la Investigación en Seguridad de la Información	Director
2015-2016	Desarrollo de Prototipos para Redes de Comunicaciones basados en Hardware Libre	Colaborador

Publicaciones, patentes, prototipos o productos (cite las más relevantes dentro de los últimos cinco años y que se encuentren alineados al proyecto de investigación)	
1.	Rodríguez-Hoyos, A., Estrada-Jiménez, J., Rebollo-Monedero, D., Parra-Arnau, J., & Forné, J. (2018). Does k-Anonymous Microaggregation Affect Machine-Learned Macrotrends?. IEEE Access.
2.	Rodríguez-Hoyos, A., Estrada-Jiménez, J., Urquiza-Aguiar, L., Parra-Arnau, J., & Forné, J. (2018, April). Digital Hyper-Transparency: Leading e-Government Against Privacy. In eDemocracy & eGovernment (ICEDEG), 2018 International Conference on (pp. 263-268). IEEE
3.	Estrada-Jiménez, J., Parra-Arnau, J., Rodríguez-Hoyos, A., & Forné, J. (2017). Online advertising: Analysis of privacy threats and protection approaches. Computer Communications, 100, 32-51.
4.	Estrada, J. A., Estrada, J. C., Rodríguez, A. F., & Tipantuña, C. J. (2015). Ecuador y la Privacidad en Internet: Una Aproximación Inicial. Revista Politécnica, 36(1), 54.
5.	Estrada-Jiménez, J. A., Rodríguez, A. F., Parra, J., & Forné, J. (2014). Evaluation of a Query-Obfuscation Mechanism for the Privacy Protection of User Profiles. Network Protocols and Algorithms, 6(2), 55-92.

Experiencia profesional, otros trabajos científicos y técnicos (cite lo más relevante o las más recientes)
Experiencia en la administración de servicios en plataformas operativas Linux y solvencia técnica en temas relacionados con Seguridad de la Información. Estuve a cargo de la gestión de proyectos relacionados con tecnologías de la información en el Ministerio de Telecomunicaciones, y administré la infraestructura tecnológica de varias entidades como: Ministerio del Interior, Sistema Nacional de Nivelación y Admisión e Interactive. He obtenido certificaciones en Linux LPI, RedHat RHCE, Cisco – CCNA y CCAI, y de Hacking Ético CEH. Colaboro con el proyecto europeo CIPSEC en la construcción de una herramienta de anonimato de logs de ciberseguridad.

PROYECTO DE INVESTIGACIÓN INTERNOS SIN FINANCIAMIENTO O AUTOGESTIONADOS

ANEXO 2 – DETALLES DE LA PROPUESTA

Investigación Básica ❑ Investigación Aplicada **X**

DEPARTAMENTO(S) Y/O INSTITUTO(S):
1. Departamento de Electrónica, Telecomunicaciones y Redes de Información
2.

LINEA(S) DE INVESTIGACIÓN:
1. Seguridad y Privacidad
2.

DISCIPLINA CIENTÍFICA ***(Marque X, solamente una opción)***	
Ciencias Naturales y Exactas;	
Ingeniería y Tecnología;	X
Ciencias Médicas;	
Ciencias Agrícolas;	
Ciencias Sociales;	
Humanidades	

OBJETIVO SOCIOECONÓMICO ***(Marque X, solamente una opción)***	
Exploración y explotación del medio terrestre;	
Ambiente;	
Exploración y Explotación del espacio;	
Transporte, telecomunicaciones y otras infraestructuras;	X
Energía;	
Producción y tecnología industrial	
Salud;	
Agricultura;	
Educación;	
Cultura, ocio, religión y medios de comunicación;	
Sistemas políticos y sociales, estructuras y procesos;	
Defensa;	
Avance general del conocimiento: I+D financiada con los Fondos Generales de Universidades (FGU);	
Avance general del conocimiento: I+D financiados con otras fuentes.	

1 **Proyecto de Investigación**

Título:
Privacidad Sintáctica Funcional: Análisis y adaptación de mecanismos de anonimato con enfoque en la preservación de utilidad de los datos

Resumen del proyecto (máximo 200 palabras)

Este proyecto aborda retos relacionados con tecnologías para protección de privacidad de individuos cuyos datos demográficos y confidenciales (microdatos), deben ser recolectados, analizados, o publicados. El espectro de impacto de estas tecnologías es muy amplio dados los crecientes volúmenes de datos que manejan los sistemas de información actualmente, e.g., sistemas de recomendación, redes sociales, o estudios socioeconómicos. Para ofrecer privacidad en este contexto, un mecanismo famoso de protección es la microagregación k-anónima que distorsiona datos para de-identificarlos.
Para obtener privacidad más "usable", planteamos extender la aplicabilidad de la tecnología de microagregación k-anónima en dos sentidos. Por un lado, proponemos estudiar y adaptar la complejidad de cómputo del algoritmo de microagregación para facilitar su aplicación en bases de datos de gran escala. Por otro lado, proponemos incorporar, en el mecanismo de protección de privacidad, adaptaciones que permitan preservar la utilidad pese a la distorsión aplicada. El ámbito de aplicación de estas propuestas es abundante, considerando la inmensa cantidad de microdatos que administran los sistemas de información (big data) actualmente y la necesidad que tienen de obtener mayor utilidad de los datos mientras ofrecen un nivel razonable de privacidad a sus usuarios.

Palabras clave (4-6):
privacidad, k-anonimato, microagregación, bases de datos, utilidad

2 Objetivos, relevancia, productos y resultados esperados de esta propuesta de investigación

Hipótesis

El algoritmo de microagregación k-anónima MDAV puede adaptarse para mejorar su desempeño en entornos de big data, manteniendo o mejorando la utilidad de los datos anonimizados.

2.1 Objetivos

2.1.1 Objetivo General

- Analizar y adaptar mecanismos de anonimización de datos, basados en privacidad sintáctica, con el fin de garantizar la usabilidad y la utilidad de los datos.

2.1.2 Objetivos Específicos

a. Estudiar el estado del arte de los mecanismos de anonimización para ofrecer privacidad sintáctica y su compromiso con la utilidad de los datos.
b. Analizar el desempeño de la microagregación k-anónima en términos de la distorsión que impone a los datos, su complejidad de cómputo, y de la utilidad de los datos resultantes.
c. Proponer mejoras computacionales en la implementación de microagregación k-anónima para facilitar su adopción en entornos de big data (microagregación a gran escala).
d. Adaptar el mecanismo de microagregación k-anónima con el fin de preservar la utilidad de los datos.

2.2 Detalle de los resultados esperados (con relación a los objetivos)

a. Estado del arte de la microagregación k-anónima con enfoque en su compromiso con la utilidad de los datos.
b. Análisis que caracterice la microagregación k-anónima de acuerdo a 3 parámetros ligados con su empleo en entornos de big data: tiempo de ejecución, distorsión sintáctica de los datos y utilidad empírica resultante.
c. Métodos que mejoren la usabilidad de la microagregación k-anónima en bases de datos de gran escala.
d. Mecanismo de microagregación k-anónima orientado a preservar la utilidad de los datos.

3 Relevancia de la propuesta de investigación y su relación con la(s) líneas de investigación

El enfoque de nuestra propuesta hacia la privacidad y los mecanismos de protección es cada vez más relevante por dos factores: la creciente disponibilidad de grandes volúmenes de datos y la permanente necesidad de extraer de ellos toda la utilidad posible. Estos fenómenos se conjugan y contraponen con la necesidad de intimidad o privacidad inherente al ser humano, provocando un compromiso que nosotros planteamos estudiar. En lo que respecta a la generación de datos a gran escala, podemos comentar varios ejemplos. El grupo de investigación IDC predijo en 2017 que el mundo estará creando 163 zetabytes de datos por año para el 2025. Esta cifra astronómica es, en parte, producto del avance tecnológico exponencial en lo referente a transmisión, almacenamiento, y procesamiento de información. Sin embargo, un factor fundamental para la generación masiva de datos es también el acceso de billones de usuarios a Internet cuyas interacciones son registradas permanentemente por varias entidades de servicio. De hecho, millones de cámaras, micrófonos y sensores desplegados en todo el mundo ya recopilan y almacenan grandes cantidades de información de los ciudadanos y sus interacciones sin necesidad de que estén en línea. Esta era en la que la información se multiplica casi sin control, la era del Big Data, trae consigo muchos beneficios basados en la explotación de esa información.

Esta explotación generalmente implica el aprovechamiento de la información para sustentar la toma de decisiones a través, por ejemplo, de la detección de tendencias o anomalías. En este sentido estratégico, la información tiene aplicaciones muy variadas en ámbitos como la medicina, el mercadeo, la política, etc. Asimismo, la gran utilidad que se puede extraer de la información ya está generando ingentes ganancias económicas en el sector privado. De hecho, dos de las empresas tecnológicas más grandes del mundo, Google

y Facebook, tienen un negocio que se basa casi enteramente en la recolección, procesamiento y publicación de información. Así, y como consecuencia de la masificación del acceso y la generación de datos, cada vez más entidades están interesadas en explotar la utilidad de sus datos, comúnmente mediante técnicas estadísticas. En este contexto en el que abunda la información digital, es inevitable encontrar o derivar información personal En la práctica, cualquier sistema de información que reciba interacción humana podría asociar información potencialmente sensible a individuos específicos debido a la gran cantidad de atributos que estos sistemas recolectan a partir de dichas interacciones. Esta información agregada, si bien es útil para orientar el servicio que ofrece una empresa, pero, al analizarla para entidades individuales (como personas), podría afectar su privacidad. Aunque los atributos de los que se obtenga información pueden parecer inocuos, al combinarse podrían servir para identificar fácilmente a una persona, razón por la cual el riesgo a la privacidad generalmente es significativo.

La relevancia de nuestra propuesta se soporta también en la creciente necesidad de la industria de proteger los datos personales que manejan las instituciones. Regulación en esa línea, cada vez más granular y estricta, se produce alrededor del mundo. Un ejemplo de ello es la GDPR [11], con jurisdicción en Europa, que está poniendo en serios aprietos a las empresas con la posibilidad de imponer ingentes multas económicas a quienes no cumplan con dicha legislación.

Nuestra propuesta es relevante pues aborda las tres dimensiones antes descritas (privacidad, utilidad y usabilidad) y resumidas en la necesidad evidente de mecanismos de protección de privacidad que modulen el nivel de distorsión necesario sobre los datos, de acuerdo a ciertos requisitos de índole empírico como son la facilidad de procesar la información a gran escala y de mantener su utilidad. En esa medida, estamos seguras de que nuestra propuesta está relacionada estrechamente con la línea de investigación Seguridad y Privacidad del Departamento de Electrónica, Telecomunicaciones, y Redes de Información. Como se detalla más adelante, el aporte de nuestra propuesta con relación a la literatura actual se traduce en evaluar no solamente el nivel de privacidad que ofrece un mecanismo de protección, sino también otras dimensiones que podrían contraponerse, como la utilidad empírica resultante de los datos protegidos y el rendimiento del mecanismo en entornos de bases de datos a gran escala. Desde luego, para ello, la propuesta considera también la adaptación del algoritmo original de protección para obtener mayor usabilidad y utilidad. Esta aproximación al amplio problema planteado inicia su alcance al estudio y adaptación de un popular algoritmo, MDAV.

4	**Productos esperados (marcar con una "X" al menos uno de los productos no señalados)**

Tipo de Producto:	**Marcar con una "X"**
a. Disertación a la Comunidad Politécnica (obligatorio) ;	X
b. Presentación de un artículo en formato de la Revista Politécnica (obligatorio)	X
c. Proyecto de Titulación;	
d. Aplicación tecnológica construida o implementada;	X
e. Patente presentada;	
f. Perfil de proyecto de mayor impacto científico, técnico, pedagógico o de innovación.	
g. Publicación científica indexada en SCIMAGO-SCOPUS WoS SCIELO Latindex Catálogo o un artículo en congreso indexado en SCOPUS.	X

5 Descripción y metodología y diseño del proyecto

5.1 Descripción, metodología y diseño del proyecto (Máximo dos carillas)

Estado del arte

En trabajos previos se ha demostrado que la combinación de unos pocos atributos de información de un individuo como fecha de nacimiento, género y código postal, conocidos como cuasi-identificadores, podría ser suficiente para identificar a dicho individuo unívocamente en una gran población [1]. Es decir, aunque se eliminen los atributos identificadores (por ejemplo, número de cédula) de una base de datos publicada, otros atributos aparentemente inofensivos, al combinarse, pondrían en riesgo la privacidad de los sujetos de dicha información pues podrían permitir la revelación de un atributo confidencial, como el estado de salud o la inclinación política.

Para controlar la privacidad en este contexto, el k-anonimato es una propiedad de una base de datos que garantiza que no se puedan reidentificar los individuos a los que se refieren los datos. Una base de datos es k-anónima si la información de cada individuo (su registro correspondiente) es la misma que la de otros k-1 individuos. Esto se puede conseguir mediante el mecanismo de microagregación [3] listado en la Fig. 1. Este mecanismo agrupa los registros de una base de datos en celdas de tamaño k (k-anónimas) y en cada celda reemplaza los cuasi-identificadores de cada registro por una misma tupla, conocida como centroide. En principio, al liberar la versión microagregada de la tabla, un atacante no podrá descubrir la identidad de un individuo a partir de un registro pues los valores de sus atributos (cuasi-identificadores) serán los mismos que los de otros k-1 individuos. Debe notarse que, comúnmente, el k-anonimato se aplica a los atributos cuasi-identificadores pues se considera que los identificadores son directamente suprimidos y que los confidenciales son necesarios para la utilidad de los datos. Gracias al k-anonimato, un atacante no podrá determinar la identidad de un participante y así tampoco asociarle su atributo confidencial.

Existen varios algoritmos de microagregación que emplean métodos heurísticos pues se trata de un problema NP-hard [2]. El algoritmo MDAV (maximum distance to average vector) [3] tiene un rendimiento muy bueno en términos de la distorsión que agrega a los datos mientras conforma celdas de tamaño fijo. Otros algoritmos con menos restricciones (como usar celdas de tamaño variable) son μ-Approx [4], SMT [5], y TFRP [6]. Otros esfuerzos abordan la microagregación reduciendo la dimensionalidad de los datos (ie., el número de atributos) mediante proyecciones en una dimensión [7], aunque presentan un riesgo mayor de privacidad.

Aunque el k-anonimato es un criterio muy útil y popular, tiene problemas relacionados con el hecho de que solo se enfoca en los atributos cuasi-identificadores. Así, no toma en cuenta la distribución resultante de los valores de los atributos confidenciales en cada celda. Si esta distribución es homogénea o si difiere mucho de la distribución en la base de datos o en la población en general, podría revelar información de los participantes de la base de datos y facilitar ataques a la privacidad como el de homogeneidad [8] o el de sesgo [9]. Para enfrentar estas deficiencias, se han planteado otros criterios complementarios como p-sensitivity [9], l-diversity [8] o t-closeness [9] que, aunque incrementan las garantías de privacidad, implican una mayor distorsión de los datos.

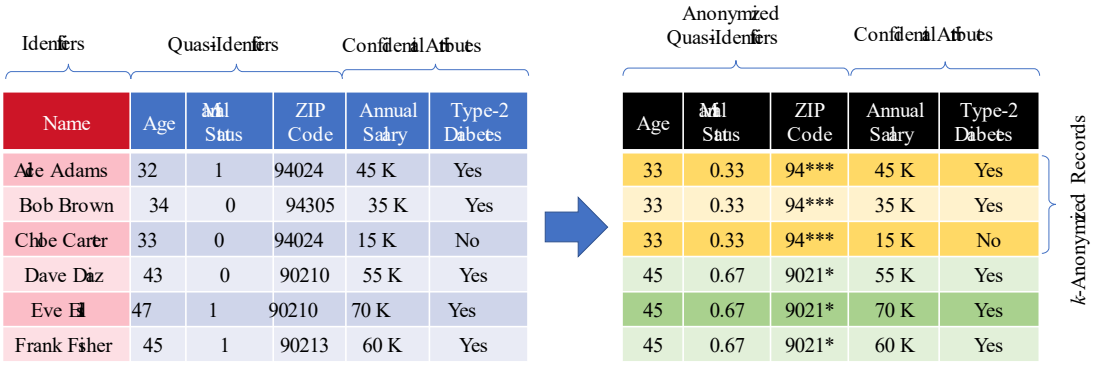

Fig. 1. Ejemplo de microagregación k-anónima de datos con k=3 que muestra información posiblemente sensible (salario anual y presencia de diabetes tipo 2) como atributos confidenciales, en relación con variables demográficas (edad, estado civil y código postal) como cuasi-identificadores. La tabla original, en el atributo de estado civil codifica casado con 1 y soltero con 0. Cuando se aplica microagregación, el atributo estado civil es anonimizado reemplazando los valores en una celda por la proporción de personas casadas en esa celda, tal como MDAV lo hará. Por ejemplo, en la primera celda (tres

primeros registros) de la tabla microagregada, el atributo estado civil corresponde a 1/3=0.33 ya que, en la tabla original una de las tres personas está casada en esa celda [12].

Además, pese a que la microagregación k-anónima puede llegar a ofrecer excelentes resultados en cuanto a distorsión de los datos [10] en comparación con otros mecanismos, este beneficio tiene un costo alto en términos de tiempo de cómputo. Esto podría hacer de la microagregación un mecanismo poco viable para aplicaciones de big data, especialmente cuando el tiempo de ejecución del algoritmo puede crecer con el cuadrado del número de registros de la base de datos.
Sin duda, el estudio del compromiso entre utilidad y privacidad es primordial para plantear criterios de privacidad y utilidad adecuados, y algoritmos más privados, útiles y aplicables, dependiendo del contexto.

Descripción del proyecto
Este proyecto pretende abordar dos dimensiones en el ámbito de la privacidad: la de las métricas y la de los métodos de implementación. En cuanto a las métricas de privacidad, nos concentraremos fundamentalmente en el k-anonimato. Con respecto a las métricas de utilidad de los datos, consideraremos el error cuadrático medio, pero evaluaremos otras que ofrezcan una caracterización empírica de la utilidad de los datos anonimizados. En el ámbito de las métricas, explorareremos métricas relacionadas con el uso de recursos de los mecanismos para implementación de k-anonimato. Finalmente, en la dimensión de métodos de implementación, investigaremos estrategias para reducir el tiempo de ejecución de los algoritmos de microagregación, así como mecanismos para preservar la utilidad de los datos anonimizados.

Metodología

Aunque hay estudios recientes relacionados con la temática planteada [13] [14], la multidimensionalidad del problema permite que se pueda abordar desde muchos frentes. Por ejemplo, en [13] se comparan, sin criterios muy claros, varios algoritmos de k-anonimato (los llaman "DataFly", "Greedy", "Samarati", "OLA", "Flash"). Nuestra propuesta se enfoca en otro algoritmo de protección de privacidad, igualmente popular, MDAV, y pretende adaptarlo para conseguir dos cosas: un mayor desempeño de éste en entornos de big data (usabilidad) y una mayor utilidad de los datos. En [14] se plantea un modelo de k-anonimato basado en el algoritmo Mondrian para reducir la pérdida de utilidad de los datos anonimizados, por lo que tampoco está relacionado con nuestra propuesta, que, como primera aproximación a la investigación en privacidad, se enfoca en el algoritmo MDAV. Al concentrarnos en este algoritmo, lo implementaremos íntegramente usando Julia y, consecuentemente, no usaremos otros frameworks como [15] con varias implementaciones ya integradas, aunque podría ser de mucha utilidad para trabajos posteriores.

Para abordar el impacto del k-anonimato en la utilidad de los datos, primero se hará una revisión del estado del arte de los mecanismos de microagregación existentes. Nos concentraremos en analizar los distintos criterios empleados para medir la privacidad, la distorsión, y, particularmente, la utilidad de los datos, luego de que estos son anonimizados. Como preámbulo para cumplir los objetivos experimentales posteriores, nos familiarizaremos con la implementación del algoritmo de microagregación MDAV, así como con los distintos tipos de conjuntos de datos que podrían servir para evaluarlo. Cabe destacar que, para esto último, nos decantaremos por el uso de Julia como herramienta de implementación de algoritmos y manipulación de datos.

Utilizando los criterios de distorsión y utilidad revisados, se analizará experimentalmente el desempeño de la microagregación k-anónima. Para ello, se comparará la distorsión y utilidad de varios datasets anonimizados con su versión original sin perturbación. Con el fin de hacer una valoración más práctica, no solo se considerará la distorsión como criterio de utilidad, sino otra métrica más empírica, como la precisión en la predicción de algún atributo. En el marco de nuestras contribuciones, *exploraremos experimentalmente la complejidad de MDAV*, en términos del tiempo de ejecución que se medirá en función del tamaño del conjunto de datos y del tamaño de las celdas.

En esa línea, con el fin de *adaptar el mecanismo de microagregación k-anónima*, estudiaremos métodos que permitan *mejorar su desempeño* en términos del tiempo de ejecución y así facilitar su adopción en entornos de bases de datos de gran escala. Dada la gran cantidad de operaciones repetitivas que se derivan de la microagregación, existen algunas opciones de optimización. Para conseguir este objetivo, analizaremos los procedimientos intermedios que conforman el algoritmo MDAV y buscaremos acelerarlos para reducir el

tiempo total de ejecución. Del mismo modo, propondremos estrategias para disminuir la cantidad de operaciones necesarias del algoritmo de microagregación mediante la reducción de la dimensionalidad de los datos; todo esto con el fin de *mejorar el desempeño del algoritmo* en términos de su uso de recursos. Se implementarán distintos métodos y se evaluarán sobre varios datasets para determinar experimentalmente su impacto.

Finalmente, ya que la usabilidad de los mecanismos de anonimato depende, en gran medida, de la utilidad resultante de los datos anonimizados, planteamos investigar métricas y adaptaciones que permitan preservar dicha utilidad mientras se ofrece un determinado nivel de privacidad. *Propondremos y evaluaremos estas estrategias en el marco de la microagregación k-anónima.*

Bibliografía

[1] Sweeney, L. (2000). Uniqueness of simple demographics in the US Population, en LIDAP-WP4. Recuperado de http://privacy.cs.cmu.edu/dataprivacy/papers/LIDAP-WP4abstract.html (Abril 2018).
[2] Oganian, A., & Domingo-Ferrer, J. (2001). On the complexity of optimal microaggregation for statistical disclosure control. Statistical Journal of the United Nations Economic Commission for Europe, 18(4), 345-353.
[3] Domingo-Ferrer, J., & Mateo-Sanz, J. M. (2002). Practical data-oriented microaggregation for statistical disclosure control. IEEE Transactions on Knowledge and data Engineering, 14(1), 189-201.
[4] Domingo-Ferrer, J., Sebé, F., & Solanas, A. (2008). A polynomial-time approximation to optimal multivariate microaggregation. Computers & Mathematics with Applications, 55(4), 714-732.
[5] Laszlo, M., & Mukherjee, S. (2005). Minimum spanning tree partitioning algorithm for microaggregation. IEEE Transactions on Knowledge and Data Engineering, 17(7), 902-911.
[6] Chang, C. C., Li, Y. C., & Huang, W. H. (2007). TFRP: An efficient microaggregation algorithm for statistical disclosure control. Journal of Systems and Software, 80(11), 1866-1878.
[7] Rebollo-Monedero, D., Forné, J., Soriano, M., & Allepuz, J. P. (2016). k-Anonymous microaggregation with preservation of statistical dependence. Information Sciences, 342, 1-23.
[8] Machanavajjhala, A., Gehrke, J., Kifer, D., & Venkitasubramaniam, M. (2006, April). l-diversity: Privacy beyond k-anonymity. In Data Engineering, 2006. ICDE'06. Proceedings of the 22nd International Conference on (pp. 24-24). IEEE.
[9] Li, N., Li, T., & Venkatasubramanian, S. (2007, April). t-closeness: Privacy beyond k-anonymity and l-diversity. In Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on (pp. 106-115). IEEE.
[10] Rebollo-Monedero, D., Forné, J., Pallarès, E., & Parra-Arnau, J. (2013). A modification of the Lloyd algorithm for k-anonymous quantization. Information Sciences, 222, 185-202.
[11] General Data Protection Regulation, Recuperado de https://eugdpr.org/ (Octubre, 2018)
[12] Rodríguez-Hoyos, A., Estrada-Jiménez, J., Rebollo-Monedero, D., Parra-Arnau, J., & Forné, J. (2018). Does k-Anonymous Microaggregation Affect Machine-Learned Macrotrends?. IEEE Access, 6, 28258-28277.
[13] Pal, D., Bhapatra, R. K., & Babu, K. S. (2017, May). Evaluation of generalization based K-anonymization algorithms. In Sensing, Signal Processing and Security (ICSSS), 2017 Third International Conference on (pp. 171-175). IEEE.
[14] Liu, K. C., Kuo, C. W., Liao, W. C., & Wang, P. C. (2018, August). Optimized Data de-Identification Using Multidimensional k-Anonymity. In 2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE) (pp. 1610-1614). IEEE.
[15] ARX - Data Anonymization Tool. Recuperado de https://arx.deidentifier.org/ (Octubre, 2018)

6 Infraestructura, equipos y fondos adicionales.

6.1 Infraestructura y equipos

- *Indicar la infraestructura y equipos **disponibles** para la ejecución del proyecto, con la ubicación actual de los mismos*

Infraestructura	**Equipos**	
	Nombre del Equipo	**Ubicación del Equipo**
Intel Core i7-7700 CPU 3.6 GHz, 16 GB RAM	Computador personal, director	

Intel Core i7-7700 CPU 3.6 GHz, 16 GB RAM	Computador personal colaborador	

6.2 Breve justificación del equipo requerido

No se solicita equipo adicional pues los algoritmos planteados para este proyecto pueden ser implementados y evaluados en los computadores personales de los colaboradores del proyecto.

6.3 Fondos Adicionales

No existen otros fondos que financien directamente este proyecto.

ESCUELA POLITÉCNICA NACIONAL

Proyecto de Investigación Interno Sin Financiamiento o Autogestionado

ANEXO 3 - CRONOGRAMA DE ACTIVIDADES DEL PROYECTO

Título del Proyecto: Privacidad Sintáctica Funcional: Análisis y adaptación de mecanismos de anonimato con enfoque en la preservación de utilidad de los datos

AÑO 1

Nº	Actividad	Mes 1				Mes 2				Mes 3				Mes 4				Mes 5				Mes 6				Mes 7				Mes 8				Mes 9				Mes 10				Mes 11				Mes 12			
		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1	**Estudiar el estado del arte de los mecanismos de anonimización para ofrecer privacidad sintáctica y su compromiso con la utilidad de los datos.**																																															
1,1	Revisar en la literatura los aportes teóricos referentes al criterio de k-anonimato para privacidad sintáctica.																																															
1,2	Explorar el estado del arte de los mecanismos para implementar k-anonimato, particularmente la microagregación k-anónima																																															
1,3	Investigar otros criterios de privacidad que complementen al de k-anonimato.																																															
1,4	Revisar en la bibliografía las principales métricas de utilidad de datos k-anónimos.																																															
1,5	Plantear una discusión sobre el compromiso entre la privacidad y la utilidad empírica de los datos.																																															
1,6	Implementar el algoritmo de microagregación k-anónima basado en las métricas estudiadas.																																															
2	**Analizar el desempeño de la microagregación k-anónima en términos de la distorsión que impone a los datos, su complejidad de cómputo, y de la utilidad de los datos resultantes.**																																															
2,1	Evaluar el impacto de la microagregación usando métricas de utilidad estándar para distintos parámetros de privacidad.																																															
2,2	Evaluar la calidad de estas métricas estándar de utilidad, comparándolas con métricas empíricas.																																															

3	**Proponer métodos que mejoren computacionalmente la implementación de microagregación k-anónima para facilitar su adopción en entornos de big data (microagregación a gran escala).**	
3,1	Explorar teórica y experimentalmente la complejidad de la microagregación k-anónima, en términos de su tiempo de ejecución.	
3,2	Plantear mejoras computacionales al algoritmo de microagregación k-anónima para reducir su tiempo de ejecución en grandes bases de datos, mejorando así la usabilidad de este algoritmo en entornos de big data.	
3,3	Evaluar estas mejoras computacionales.	
4	**Adaptar el mecanismo de microagregación k-anónima con el fin de preservar la utilidad de los datos.**	
4,1	Examinar las deficiencias del criterio de k-anonimato en relación a la utilidad resultante de los datos anonimizados.	
4,2	Proponer estrategias para incrementar la utilidad de los datos anonimizados con microagregación k-anónima	
4,3	Evaluar experimentalmente estas estrategias.	