UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

Manteniendo la privacidad en los censos nacionales

INFORME FINAL DE CC6907 PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL EN COMPUTACIÓN

Arturo Kullmer

MODALIDAD:
Memoria

PROFESOR GUÍA:
Federico Olmedo
PROFESOR GUÍA 2:
Matías Toro

SANTIAGO DE CHILE
2024

1. Introducción

En el contexto actual, los datos son recogidos, almacenados y estudiados a gran escala, la privacidad de los datos se ha convertido en una gran preocupación, donde las instituciones públicas, que manejan datos altamente sensibles, deberían resguardar la privacidad de las personas.

El reciente crecimiento en el poder computacional facilita cada vez más la violación de la privacidad de los individuos. Esto hace necesario invertir esfuerzos en proteger la privacidad de los datos para prevenir vulnerabilidades.

En el ámbito de la computación, la privacidad de los datos se refiere a la protección de la información sensible de distintas entidades, incluso cuando estos datos tienen disponibilidad pública. Esto hace que la privacidad sea especialmente desafiante, ya que cualquier persona podría acceder a los datos y extraer conocimiento. Así, se genera una obligación en asegurar que los datos mantengan la confidencialidad necesaria para proteger la identidad y la información personal de los individuos involucrados.

Esto puede parecer similar a la seguridad de los datos, ya que ambos tienen un fin en común: mantener seguros los datos privados y sensibles de las personas y organizaciones. Sin embargo, la seguridad de los datos se centra en la protección de la integridad, disponibilidad y confidencialidad de los datos, sean o no publicados, a través de medidas de seguridad informáticas. En cambio, la privacidad busca proteger la información personal, asegurando que incluso los datos que sí son públicos mantengan la privacidad adecuada al momento de hacerlos disponibles.

Hoy en día, existe un gran interés en publicar información, ya que esto aumenta la transparencia de las organizaciones, lo que suele incrementar la confianza y la imagen pública de estas. Es por esta razón que, en el año 2008, en Chile se promulgó la Ley N° 20.285 sobre Acceso a la Información Pública, mejor conocida como la Ley de Transparencia [14]. Esta ley obliga a las entidades públicas a proporcionar información cuando sea solicitada, siempre y cuando no se comprometa la seguridad y privacidad de las personas o el Estado. Además, la ley también establece la llamada Transparencia Activa, que obliga a los organismos públicos a publicar datos de manera proactiva, con el objetivo de mejorar la transparencia del Estado.

Esto hace que, para poder publicar cualquier información, sea imprescindible realizar esfuerzos para mejorar la privacidad. De hecho, es una obligación, ya que la Ley N° 19.628 sobre Protección de la Vida Privada, mejor conocida como la Ley de Protección de Datos Personales, responsabiliza a estos mismos organismos de garantizar la privacidad de los datos personales [13].

Es en este contexto donde el Instituto Nacional de Estadísticas (INE) realiza censos de población y vivienda aproximadamente cada 10 años, cuyos resultados deben transparentarse para que puedan ser utilizados por cualquier persona. Esto implica que el INE tiene una responsabilidad legal y ética de privatizar correctamente los datos que se publican.

Actualmente, en el año 2024, se está llevando a cabo el Censo de Población y Vivienda. Se espera que los primeros resultados estén disponibles en 2025. Los últimos datos censales

disponibles públicamente son los microdatos del censo del año 2017.

El INE, para cumplir con la Ley de Transparencia y la Ley de Protección de Datos Personales, utilizó técnicas de anonimización para proteger los datos del censo de 2017. Sin embargo, no hay certeza de que los esfuerzos realizados por esta institución sean suficientes para asegurar que los datos sean efectivamente privados, haciendo que sea de suma importancia que se investigue y evalue el nivel de privacidad de estos.

Un mecanismo con métricas más claras respecto a la privacidad de los datos censales puede otorgar gran valor al INE y a las personas del país, dado que existen importantes implicaciones legales y éticas de por medio. Además, si se detectan vulnerabilidades en la privacidad de estos datos, este mecanismo no solo sería necesario, sino imprescindible. Por último, en el contexto del censo de 2024, una solución de este tipo podría marcar un precedente para futuras publicaciones del INE y otras instituciones estatales.

2. Situación Actual

En la actualidad, existen dos principales modelos para privatizar los datos. Por un lado, está la anonimización, que es lo que actualmente está utilizando el INE para publicar microdatos del censo. Por otro lado, está la privacidad diferencial, la cual fue utilizada en Estados Unidos para privatizar los datos a publicar de su censo en 2020.

La anonimización consiste en la conversión de los datos personales para que estos no puedan ser usados para identificar a un individuo o información sensible asociada a este [7]. Las conversiones más usuales son la supresión y la generalización de datos. Por ejemplo, el censo de Chile de 2017 no publica los nombres ni apellidos de las personas. Tampoco muestra la dirección de dónde viven, pero sí el bloque censal al cual están sujetos. De esta manera, se suprimen los nombres y se generalizan las direcciones.

Las decisiones para determinar qué datos anonimizar y cuáles no, además de cuánto se deben generalizar, están sujetas al supuesto de que un atacante puede cruzar los datos con información auxiliar que tiene valores en común con los datos originales. Esto lleva al problema de que es difícil encontrar un equilibrio en cuánto se debe anonimizar, pues a priori, la información que tiene un atacante es desconocida. Por esta razón, no es fácil tener certezas sobre la privacidad, ya que es sencillo hacer un supuesto erróneo. Existen varios casos donde se ha vulnerado la información de personas. Por ejemplo, en los datos del Servel del plebiscito constitucional de 2020, Matías Toro logró encontrar que hay 65.532 personas que son susceptibles a ataques de asociación [17]. También, cuando personalmente rendí el curso de privacidad de datos, logré reidentificar hasta 20 personas de los datos públicos del MINEDUC.

Por otro lado, durante la planificación de la privatización de los datos del censo de Estados Unidos de 2020, se identificó que haber utilizado anonimización para privatizar los datos del censo de 2010 tuvo importantes fallas. Se logró reconstruir los datos personales de 144 millones de personas. Además, utilizando datos comercialmente disponibles, se alcanzó a identificar hasta 52 millones de individuos [8]. Esto llevó a que, para los datos del censo de 2020, se utilizara privacidad diferencial.

La privacidad diferencial se basa en el principio de denegación plausible, que implica que un atacante no pueda determinar con certeza si un individuo en particular está presente o no en los datos, incluso si posee información adicional. Esto se logra perturbando los resultados al agregar ruido aleatorio controlado. Tal perturbación genera incertidumbre respecto al valor real de los datos y dificulta que un atacante los vulnere. Es crucial regular cuánto ruido se agrega, ya que si es muy poco, se expone más información, pero si es excesivo, se pierde utilidad en los datos [4]. De esta manera, las garantías son distintas a las que ofrece la anonimización, pues no se supone cual información es la que posee un posible atacante.

Para encontrar un equilibrio entre la utilidad y la privacidad, se utiliza un "presupuesto de privacidad" que generalmente se denota por un ε. Tal valor regula cuánto ruido se le agrega a los datos; un valor muy alto significa que hay poca privacidad y mucha utilidad, mientras que un valor muy bajo implica mucha privacidad pero poca utilidad. En ese sentido, no es trivial encontrar cuál es el mejor valor de ε, y es una de las principales decisiones que se deben tomar a la hora de utilizar privacidad diferencial.

En Chile, existen muy pocas iniciativas para utilizar privacidad diferencial en la publicación de datos, y hasta el momento, no se han observado implementaciones significativas en instituciones públicas. Además, considerando el contexto actual en que se está desarrollando el Censo de Población y Vivienda 2024, y el hecho de que se ha demostrado que los datos publicados por el Servel y por ministerios como el MINEDUC tienen problemas de privacidad, es una señal de que es necesario revisar cómo se están privatizando los datos censales y evaluar cómo puede mejorarse este proceso con la privacidad de datos.

3. Trabajo relacionado

La experiencia internacional en la privacidad de los datos censales ha trabajado con diferentes técnicas. Por ejemplo, en varios países de Europa se ha utilizado el intercambio de datos y/o el llamado "Cell Key Method" (CKM) para tal proceso.

El intercambio de datos consiste en encontrar qué grupos son los más susceptibles a ataques, generalmente mediante cálculos de k-anonimato y l-diversidad, para luego cambiar algunos datos pertenecientes a ese grupo para que logren mezclarse con otros, protegiendo así su privacidad. Esta es una de las técnicas que actualmente usa el INE [9], y que también es recomendada por el Centro de Excelencia para el Control de la Divulgación de Estadísticas de la Unión Europea [5]. Existen muchos tipos de intercambio de datos, como "Rank Swapping" [3]. En general, todas se basan en intercambiar ciertos datos en específico con algún criterio que ayude a preservar las propiedades estadísticas de los datos.

Por otro lado, CKM tiene un elemento en común con la privacidad diferencial, puesto que también se basa en agregar ruido, aunque especificamene a la frecuencia de los datos. CKM define una llave (key) para cada celda, la cual es luego usada como semilla para ver cuánto ruido agregar. Esto hace que tal llave pueda ser utilizada para generar otras tablas para que los mismos datos contribuyan de la misma manera a través de ambos conjuntos de datos. Esto último es lo que hace atractivo al método, ya que la consistencia es fundamental para la publicación de datos.

Junto a lo anterior, el método original considera además otras propiedades interesantes para la publicación de datos censales, ya que se preocupa de que no se generen valores negativos, que el ruido esté centrado en cero, que el ruido agregado no sea mayor que el valor perturbado cuando este es suficientemente pequeño y también que se mantenga una varianza fija entre dos celdas perturbadas [6].

Estos últimos dos métodos son utilizados por la Comisión Económica de las Naciones Unidas en Europa (UNECE) para la publicación de datos censales, es más, la organización "Statistic Netherlands", que trabaja en conjunto con la UNECE, provee los software llamados "μ-Argus" y "τ-Argus" para facilitar el proceso de privatización de datos, en particular para la publicación de microdatos y datos tabulados respectivamente [16]. Sin embargo, de manera similar al INE en Chile, tales implementaciones utilizan principalmente técnicas de supresión, intercambio, agregación y redondeo de datos. En lo que respecta al uso de CKM, este tampoco otorga garantías formales como las que ofrece la privacidad diferencial, ya que no considera conceptos como el presupuesto de privacidad (ε) y la sensibilidad de las consultas.

Ahora, como se mencionó anteriormente, en el caso de Estados Unidos, el "Census Bureau" u Oficina Censal utilizó privacidad diferencial para privatizar la información censal. En particular, se desarrolló el algoritmo "TopDown" para tal propósito.

El algoritmo se basa en aprovechar la jerarquía del árbol geográfico del país para aplicar privacidad diferencial en cada nivel del árbol. TopDown empieza por el punto más alto, es decir, la raíz del árbol que representa los datos a nivel país, y luego va bajando por el árbol añadiendo ruido en cada nivel hasta llegar a las hojas, que en este caso serían los bloques censales. Este enfoque permite que, tras procesar un nodo del árbol geográfico, se pueda utilizar el resultado para procesar a los hijos de tal nodo, permitiendo así que se mantenga la consistencia entre los datos a lo largo de todo el árbol. Además, en cada nivel del árbol, TopDown se preocupa de preservar una serie de invariantes. Por ejemplo, que la cantidad de población a nivel estatal se mantenga constante.

El proceso de satisfacer los invariantes y garantizar que los nodos hijos sean consistentes con el nodo padre se denomina fase de estimación o posprocesamiento. Esto se separa en un proceso aparte ya que resulta computacionalmente muy costoso encontrar una solución que cumpla con todos los invariantes en una única ejecución [1]. Por ello, el algoritmo añade ruido en cada nivel sin considerar estas restricciones inicialmente, para luego formular un problema de optimización que encuentre una solución que satisfaga todas las restricciones, y que además, sea similar a la solución obtenida originalmente que no tenía imposición de condiciones. Este punto se detalla con mayor profundidad en la sección de trabajo adelantado.

Ahora, es muy importante satisfacer los invariantes y restricciones, esto debido a que la cantidad de escaños en la "House of Representatives" para cada estado se determina por su población [19], haciendo así que la transparencia de este dato poblacional sea de suma importancia. Además, debido al Título 13 del Código de los Estados Unidos, la oficina censal esta obligada a proveer estos datos para facilitar la toma de desiciones e investigación [18].

Tomando en cuenta lo anterior, el funcionamiento del censo en Chile tiene semejanzas con el de Estados Unidos, ya que el INE también esta en obligación de publicar los datos por la Ley de Transparencia y el territorio chileno forma un árbol geográfico que tiene componentes

similares a Estados Unidos. La raíz sigue siendo el nivel nacional y las hojas los bloques censales. También es importante la presencia de transparencia en la población de algunos niveles del árbol. Esto debido a que la cantidad de escaños de algunos roles políticos se determinan en base a estos datos. Por ejemplo, como se establece en la ley N° 20.840 respecto a la cantidad de diputados; *"Los 155 escaños se distribuirán proporcionalmente entre los 28 distritos en consideración a la población de cada uno de ellos, en base a los datos proporcionados por el último censo oficial de la población realizado por el Instituto Nacional de Estadísticas"* [15]. Destacando nuevamente la importancia del censo y el rol político que alcanza a ocupar.

4. Objetivos

Objetivo General

El objetivo de esta memoria es mejorar la protección de la privacidad de las personas en los datos del Censo de Población y Vivienda de Chile del año 2017 mediante la evaluación y estudio de los métodos formales de privatización de datos utilizados.

Objetivos Específicos

1. Evaluar los riesgos de vulneración de la privacidad asociados con los métodos de anonimización empleados por el Instituto Nacional de Estadísticas (INE) en la publicación de los datos del censo de 2017.
2. Estudiar la experiencia internacional en la privacidad de datos aplicado a datos censales.
3. Diseñar un método de privatización de datos que incremente la protección de la privacidad de los individuos del censo nacional de 2017, manteniendo un nivel de utilidad relevante.
4. Implementar y aplicar el método de privatización de datos diseñado a los datos del Censo de Población y Vivienda de Chile 2017.
5. Validar que el método de privatización efectivamente incremente la privacidad de los datos censales y que mantenga un nivel de utilidad adecuado.

Evaluación

Para evaluar la privacidad de los datos del trabajo a efectuar se realizarán ataques utilizando datos públicamente disponibles. Los resultados de dicho ataque podrán posteriormente ser comparados con un ataque similar a los datos censales que están publicados actualmente. Después, en caso de utilizar privacidad diferencial, también se podrá comparar el presupuesto de privacidad (ε) con el utilizado en otras publicaciones de datos.

Luego, para evaluar la utilidad de los datos, se compararán resultados de consultas clave a los datos actuales, como distribuciones de población, escolaridad y nacionalidad, todo esto a distintos niveles geográficos. En específico, se emplearán medidas de error como el error absoluto y el error cuadrático medio.

5. Solución Propuesta

En este trabajo de memoria se propone evaluar la privacidad de los microdatos del Censo de Población y Vivienda de Chile de 2017, publicados por el Instituto Nacional de Estadísticas (INE), y aplicar privacidad diferencial para mejorar su protección.

Para la evaluación de los datos se utilizará Python como lenguaje de programación, junto con la librería Pandas para el manejo de los datos. Adicionalmente se utilizará Matplotlib y Numpy para realizar visualizaciones y operaciones matematicas más complejas.

En primer lugar, se estudiarán las restricciones y técnicas de publicación de datos que implementó el INE para el censo. En este punto, se analizarán la Ley de Protección de la Vida Privada y la Ley de Transparencia, además de las técnicas de anonimización utilizadas por el INE.

A continuación, se estudiarán los datos publicados, identificando cuasi-identificadores y atributos sensibles. También se emplearán conceptos como el de k-anonimato y l-diversidad para medir el nivel de privacidad actual de los datos.

Posteriormente, se buscarán datos auxiliares de personas pertenecientes a grupos más pequeños en el censo, como extranjeros o personas que se identifiquen con un pueblo originario, ya que las minorías tienden a tener peores garantías de privacidad.

Con la información recabada, se procederá a atacar los datos mediante asociaciones de registro con los datos auxiliares conseguidos, para intentar reidentificar el mayor número posible de individuos. Esto servirá para demostrar las vulnerabilidades actuales en la privacidad de los datos censales.

Una vez concluido este proceso, se privatizarán los datos utilizando privacidad diferencial, tomando como caso de estudio el censo de Estados Unidos de 2020. En particular, se investigará el algoritmo denominado "TopDown", que fue utilizado para la privatización de los datos de dicho censo.

El objetivo de estudiar el censo de Estados Unidos es adaptar el algoritmo TopDown a la geografía chilena, aprendiendo de la experiencia extranjera sobre cómo aplicar privacidad diferencial a un conjunto de datos tan extenso como el censo. Esto contribuirá a tomar decisiones más informadas que reduzcan la pérdida de utilidad en los datos.

Es importante señalar que, a lo largo del desarrollo, se ajustará el presupuesto de privacidad (ε) para lograr un equilibrio entre privacidad y utilidad. En este punto se definirán métricas más precisas sobre la utilidad de los datos, con el fin de basar las decisiones respecto al valor de ε en dichas métricas.

Finalmente, se evaluarán las diferencias en términos de utilidad y privacidad entre los datos publicados por el INE y los datos privatizados con privacidad diferencial, utilizando métricas como el error absoluto y el error cuadrático medio. Estos errores medirán las diferencias en las respuestas a consultas realizadas a los datos, como por ejemplo, la población extranjera en una comuna específica.

6. Plan de Trabajo

El plan para el desarrollo del trabajo se enfocará en tres fases clave: estudio, desarrollo y validación, cada una de estas etapas está diseñada para abordar aspectos más específicos. Cabe notar que ya se ha adelantado trabajo durante este semestre, por lo que en la fase de estudio hay aspectos que no se incluirán. A continuación, se explicará que se hará en cada fase.

1. **Fase de Estudio:**
   (a) Cálculo de l-diversidad y k-anonimato para los microdatos del censo de Chile 2017.
   (b) Búsqueda de datos públicos con lo que cruzar los microdatos censales y ataques de asociación de registros para evaluar el nivel de privacidad actual.
   (c) Profundización de la investigación acerca de TopDown y de cómo realizar una adaptación a los microdatos del censo chileno.
2. **Fase de Desarrollo:**
   (a) Diseño de consultas clave para hacer a los datos del INE, que también serán usadas para aplicar TopDown a los datos nacionales.
   (b) Elección de presupuestos de privacidad ρ a utilizar en las consultas previamente diseñadas.
   (c) Diseño, implementación y ejecución del algoritmo, basado en TopDown, para privatizar los datos nacionales.
3. **Fase de Validación:**
   (a) Evaluación de la utilidad y privacidad de los datos obtenidos con la ejecución del algoritmo. Puede que sea necesario reconsiderar las decisiones tomadas en la implementación.
   (b) Documentación del proceso efectuado y los resultados del proyecto.
   (c) Escritura del informe final y preparación de la defensa del trabajo de título.

En la figura 1 se expone una carta Gantt con la duración estimada de cada fase y punto de la memoria para el semestre de Otoño 2025.

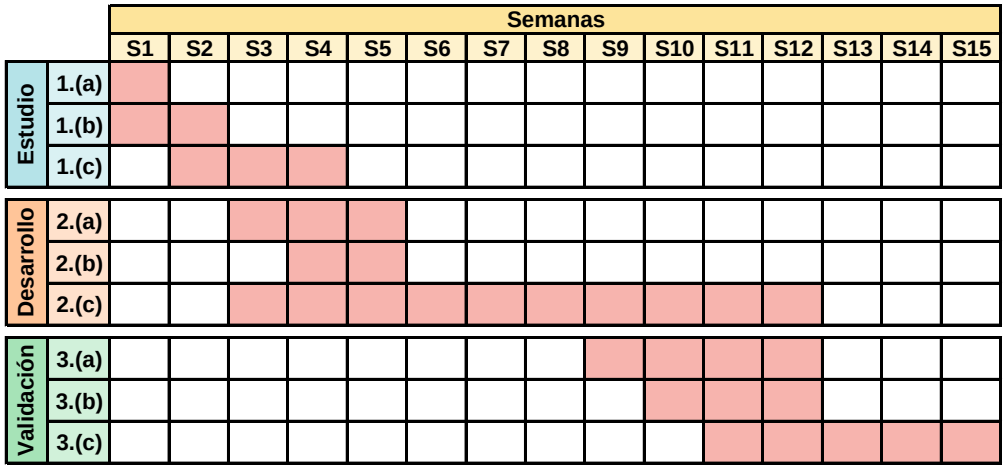

Figura 1: Carta Gantt para el semestre Otoño 2025

7. Trabajo adelantado

En lo que respecta al trabajo realizado durante el semestre del curso de "Introducción al Trabajo de Título" (CC6907), este se enfocó principalmente en investigar qué estrategias utilizó el INE para la publicación de los datos del censo de 2017, junto con cuales características tienen los datos, más específicamente, qué técnicas de privatización de datos utilizan y cómo se etiquetan los datos entre cuasi-identificadores, datos sensibles y datos no sensibles. También se indagó en qué restricciones legislativas tiene tal institución para publicar información.

Por otro lado, también se investigó en mayor profundidad el censo de Estados Unidos del año 2020, en particular, acerca de cómo utilizaron el algoritmo "TopDown" para aplicar privacidad diferencial a los datos y cómo funciona tal algoritmo. De manera similar al INE, también se revisó qué restricciones o condiciones tiene su oficina censal para publicar tales datos con el objetivo de poder comparar el proceso chileno con el estadounidense.

7.1. Censo de Población y Viviendas de Chile 2017

El Censo de Población y Viviendas de Chile de 2017 recabó información de 17.574.003 personas y 6.499.574 viviendas [10]. Estos datos son clave para la planificación de políticas públicas y la toma de decisiones, pero su publicación debe equilibrar las obligaciones de transparencia con la protección de la privacidad establecida por la Ley de Protección de Datos Personales. A continuación, se analizarán las preguntas realizadas y datos de este censo.

En los microdatos publicados, además de los datos asociados a la vivienda y población, se añade la información geográfica acerca de donde se censó tal persona en particular. Tal valor llega hasta un nivel de agregación de la zona o localidad a la cual pertenece la vivienda censada. No se da el valor específico de la dirección de la vivienda para proteger la privacidad. En la figura 2 se puede apreciar cómo el INE divide el territorio en divisiones más granulares.

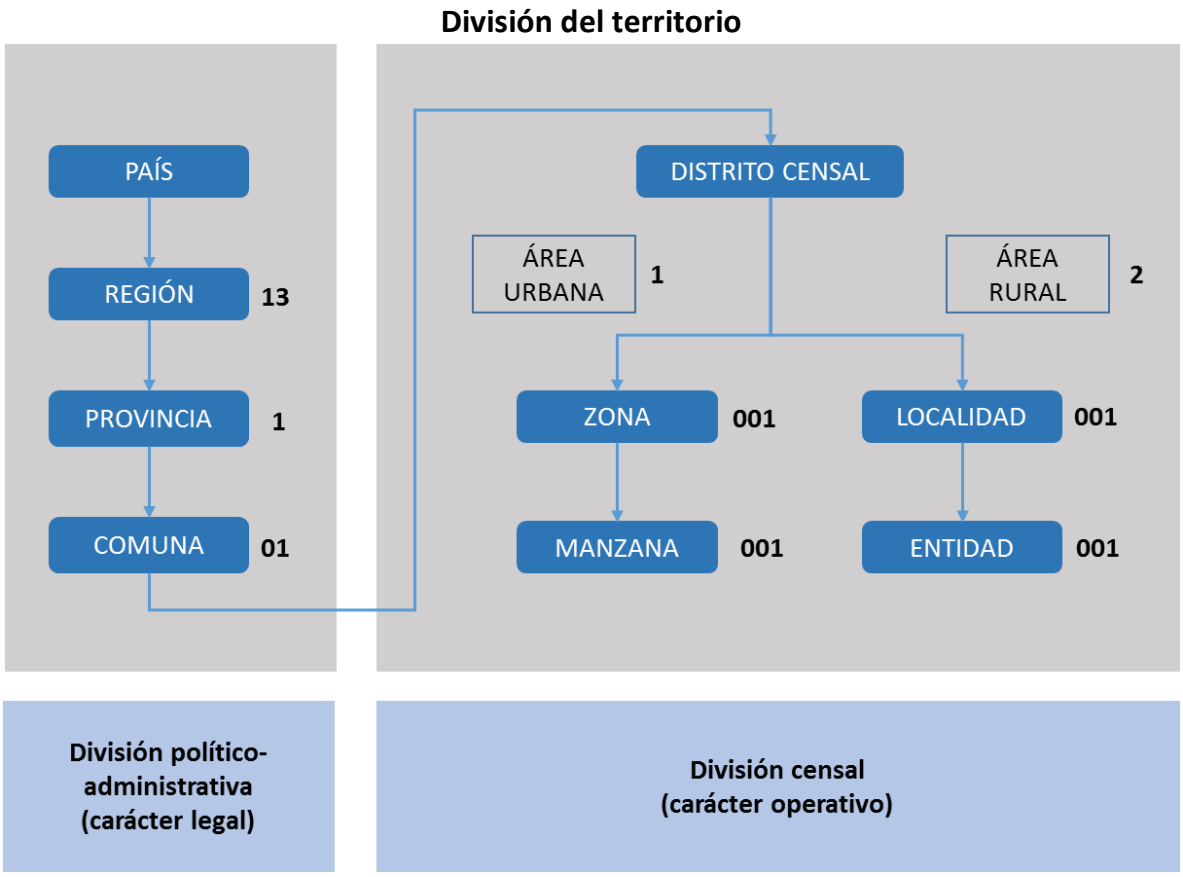

Figura 2: División del territorio por el INE para el censo [11]

El censo de personas incluye 15 preguntas principales y 4 subpreguntas, sumando 19 en total. Por su parte, el censo de viviendas contiene 6 preguntas, de las cuales 2 tienen subpreguntas. Este volumen de información, aunque útil para diversos fines, aumenta las oportunidades para que un atacante pueda vincular los datos con información auxiliar e identificar a una persona.

Para garantizar la privacidad de los datos recopilados, es importante clasificar la información según su nivel de sensibilidad:

- **Cuasi-identificadores:** Datos que individualmente no identifican a una entidad, pero que de manera colectiva pueden hacerlo, por ejemplo, la comuna de residencia, género y edad.
- **Datos sensibles:** Información cuya divulgación podría generar discriminación o perjuicios a una entidad, como la religión o el estado de salud.
- **Datos no sensibles:** Información general que no está asociada a riesgos de privacidad, como por ejemplo el material de un suelo.

Ahora, a continuación se listarán cada una de las preguntas que son incluidas en los microdatos con una descripción de a que se refiere si es necesario. También se define su

respectiva clasificación y se incluye la información geográfica. En otras palabras, se analizarán y clasificarán los microdatos.

Clasificación de microdatos del censo de viviendas

Las columnas en los microdatos están escritas de manera abreviada, se explicitará tal notación para facilitar la comprensión de los microdatos a futuro.

- Identificación geográfica hasta el nivel de localidad (**ID_ZONA_LOC**). Se considera como **cuasi-identificador**.
- Pregunta 1 (**P01**). Indica el tipo de vivienda, es decir, si es una casa, un departamento, una vivienda móvil, entre otras opciones. Se considera como **cuasi-identificador**.
- Pregunta 2 (**P02**). Indica si la vivienda está ocupada o desocupada, explicitando cuando está desocupada si es por que los moradores están ausentes, porque esta en venta o porque es una vivienda de temporada o vacaciones. Este dato se considera como **sensible** porque una persona podría aprovecharse de esta información para ver si una vivienda esta vacía para ejecutar un delito, como el de un hurto.
- Pregunta 3 (**P03A, P03B y P03C**). Indica el material con el cual estan hechas las paredes, la cubierta del techo y el piso respectivamente. Se considera como datos **no sensible**.
- Pregunta 4 (**P04**). Indica la cantidad de habitaciones que se usan como dormitorio. Se considera como **no sensibles**.
- Pregunta 5 (**P05**). Indica de donde proviene principalmente el agua que se usa en la vivienda. Se considera como **no sensible**.
- Pregunta 6 (**CANT_PER**). Identifica la cantidad de personas que pasaron la noche en esa vivienda en particular el día que fue censada. Esta pregunta se subdivide en 4, pero en los microdatos únicamente se indica la cantidad de persona. Se considera como dato **no sensible**.

De lo anterior, se puede notar que no existen muchos datos que faciliten hacer un ataque de asociación de registro, sin embargo, al haber un dato sensible es importante preocuparse de que efectivamente no se pueda realizar esto.

También cabe notar que si una vivienda está desocupada, esta no puede ser censada, pues no habría una persona capaz de responder las preguntas. El cuestionario del censo considera esta situación y no son respondidas las preguntas posteriores a la 2 si esto sucede. La figura 3 muestra el diagrama de flujo que sigue el censo de viviendas.

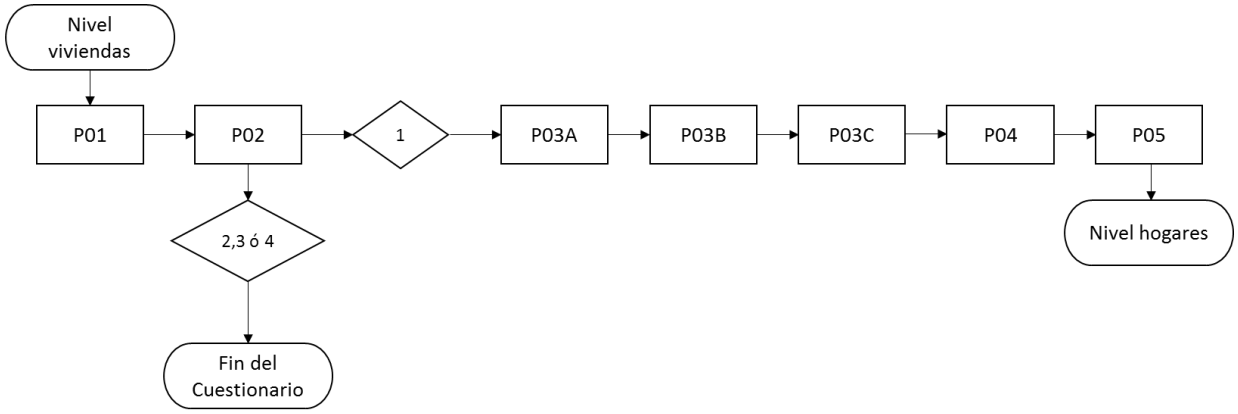

Figura 3: Diagrama de flujo del Censo de Vivienda [11]

Clasificación de microdatos del censo de población

De la misma manera que para el censo de viviendas, se incluirá la notación utilizada en los microdatos y la información geográfica presente en estos.

- Identificación geográfica hasta el nivel de localidad (**ID_ZONA_LOC**). Se considera como **cuasi-identificador**.
- Pregunta 7 (**P07**). Indica la relación de parentesco que se tiene con el jefe de hogar, por ejemplo, hijo, hermano, pareja, entre otros. Se considera como dato **no sensible**.
- Pregunta 8 (**P08**). Indica el sexo de la persona censada. Se considera como **cuasi-identificador**.
- Pregunta 9 (**P09**). Indica la edad de la persona, usa como cota superior la edad 100, es decir si una persona tiene más de 100 años se establece como si tuviese 100. Se considera como **cuasi-identificador**.
- Pregunta 10 (**P10, P10COMUNA y P10PAIS**). Indica si la persona reside habitualmente en la comuna en la fue censada, en caso de que no sea así, se pregunta en que otra comuna o país habita normalmente. Se considera como dato **cuasi-identificador**.
- Pregunta 11 (**P11, P11COMUNA y P11PAIS**). Indica si en abril de 2012 habitaba en la misma comuna que la pregunta anterior o en otra comuna o país. Se considera como **no sensible**.
- Pregunta 12 (**P12, P12COMUNA, P12PAIS, P12_LLEGADA y P12_TRAMO**). Indica en qué comuna o país habitaba la madre de la persona censada en el momento de su nacimiento. En caso de que haya nacido en otro país se pregunta que año o rango de años llego a Chile. También se considera como datos **no sensibles**.
- Pregunta 13 (**P13**). Indica si actualmente la persona censada asiste a educación formal. Se considera como dato **no sensible**.
- Pregunta 14 (**P14**). Indica el curso o año más alto aprobado considerando la respuesta de la pregunta anterior. Se considera como dato **no sensible**.
- Pregunta 15 (**P15 y P15A**). Especifica que educación formal se refiere la persona censada en los puntos anteriores. Por ejemplo, que tipo de educación media (Científico-

Humanista, Técnica profesional, etc...) o que tipo de educación superior (Técnica, Profesional, Magister o Doctorado). La pregunta P15A responde si terminó el nivel de educación declarado. Se considera como dato **sensible** puesto que una persona podría utilizar este dato, por ejemplo, para discriminar a que personas considerar en un proceso de contratación.

- Pregunta 16 (**P16, P16A y P16_OTRO**). Indica si la persona censada se considera o no perteneciente a un pueblo originario y especifica a cual en caso afirmativo. Esto se podría considerar como un **dato sensible**, pues en algunas partes del país existen problemáticas sociales, territoriales, militares y sesgos asociados a pueblos indígenas en especifico. Así, una persona que se considera parte de estos grupos, pero que quiere evitar ser relacionada con estas problemáticas, es probable que quiera mantener de manera privada tal información.
- Pregunta 17 (**P17**). Indica si trabajó o no la semana pasada. Se especifica según corresponda en hasta 6 categorías. Este dato se considera como **no sensible**.
- Pregunta 18 (**P18**). Indica el rubro al cual se dedica la empresa, institución o actividad independiente en la cuál trabajó la persona censada. Se considera como **no sensible**.
- Pregunta 19 (**P19**). Indica la cantidad de hijos nacidos vivos que ha tenido una mujer mayor a 15 años. Se considera como dato **sensible** junto a las siguientes 3 preguntas.
- Pregunta 20 (**P20**). Indica la cantidad de esos hijos que, en el momento de ser censada la madre, siguen estando vivos.
- Pregunta 21 (**P21M y P21A**). Indica el mes y año que nació el último hija o hijo de la madre censada. Estas 3 preguntas permiten inferir cuantos hijos de una madre han fallecido, lo cual es muy probable que quiera ser privatizado por una madre ya que la muerte de un hijo es un evento personal y doloroso, considerado como información altamente sensible desde el punto de vista emocional y social. Esto podría dar lugar a estigmatización o discriminación en ciertos contextos.

Considerando la clasificación anterior, se puede notar que para el censo de personas existe una mayor cantidad de datos para una posible reidentificación por parte de un atacante. Además, también existe una mayor presencia de datos sensibles. Esto nuevamente refuerza lo importante que es implementar un mecanismo de privatización de estos datos lo suficientemente fuerte para evitar exponer los datos de las personas del censo.

Al igual que el censo de viviendas, hay preguntas que no son respondidas por todas las personas. El flujo que sigue este cuestionario es bastante más complejo, por ejemplo, las preguntas posteriores a la 16 no son respondidas si la persona censada es menor a 15 años. En la figura 4 se puede apreciar el diagrama de flujo que sigue el cuestionario del censo de personas.

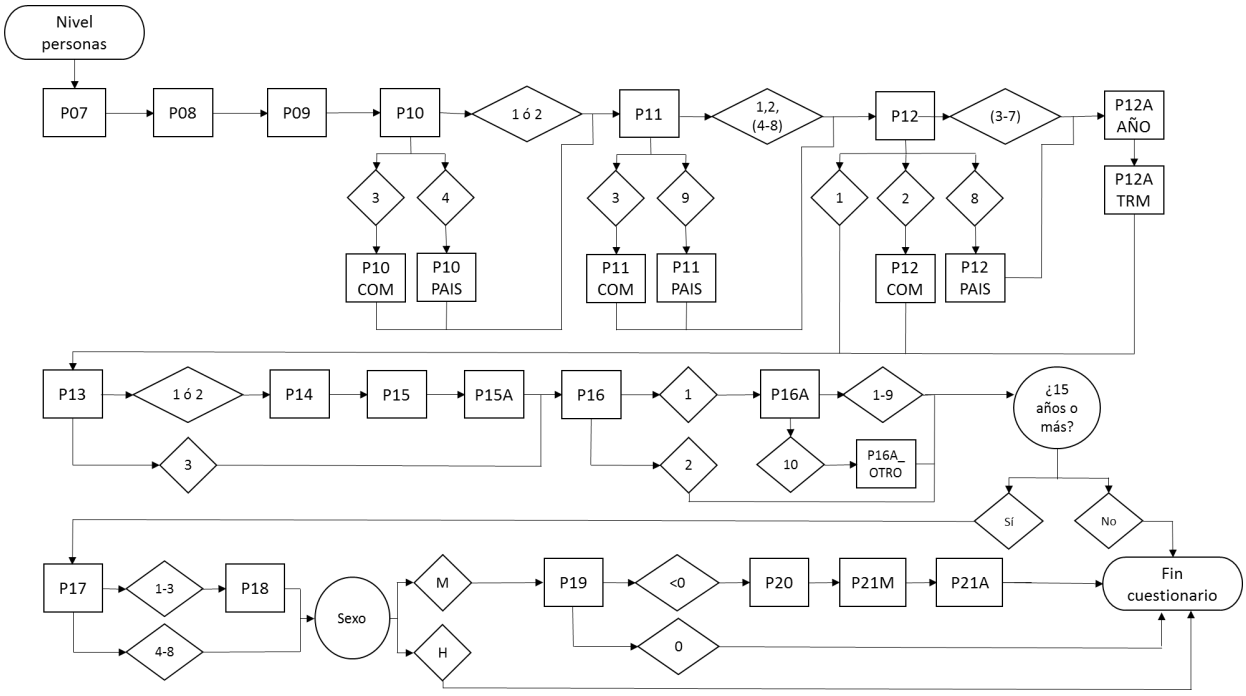

Figura 4: Diagrama de flujo del Censo de Personas [11]

7.2. Métodos de privatización usados en el censo

El Instituto Nacional de Estadísticas (INE) utiliza diversas técnicas para proteger la confidencialidad de los datos al publicarlos. Sin embargo, la documentación oficial deja de manera ambigua acerca de específicamente cuál es la totalidad de métodos que utilizaron. En particular, respecto a la indeterminación de las bases de datos el INE declara:

> *"Para realizar la indeterminación se revisó experiencia internacional, con el fin de tener una mirada general del panorama en torno a la protección de los datos de la población. En estas experiencias se constataron altas restricciones en torno a la protección de los datos, siendo los principales métodos de indeterminación el intercambio de datos, la recodificación de variables y categorías, la restricción del nivel de información geográfica, la entrega de valores determinables como "no disponibles", el redondeo de datos con límites máximos y mínimos, la entrega solo de muestras de la población y el intercambio aleatorio para muestras pequeñas de hogares similares en áreas cercanas, entre otros"* [11].

A pesar de lo anterior, el INE sí menciona algunas de las decisiones que tomaron para la indeterminación. Por ejemplo, la decisión de quedarse con un nivel de agregación geográfico a nivel de zona o localidad en vez de manzana o entidad. También se indica que en las manzanas o entidades que presentan 3 personas o menos, se suprimió información relacionada al sexo, edad, pertenencia a pueblos indígenas y migración. Además, se menciona que cuando las manzanas o entidades no cumplen con la restricción de que tengan más de 3 viviendas, entonces se agrupan múltiples manzanas/entidades para lograr una mejor privacidad.

Por último, el INE menciona también que ciertos datos son agrupados debido a que se forman grupos muy pequeños. Esto incluye a las personas mayores a 100 años, a los países de residencia y pueblos originarios con muy poca presencia y el año de llegada al país.

7.3. Algoritmo TopDown

Como se mencionó en la sección de trabajo relacionado, se investigó en mayor profundidad el algoritmo TopDown. Esto debido a que, gracias a los múltiples componentes en común que tiene el censo de Estados Unidos con el de Chile, adaptar tal algoritmo al censo nacional es una alternativa excelente para hacer un mecanismo de privatización con mayores garantías formales debido a la privacidad diferencial. En la figura 5 se puede ver el funcionamiento a grandes rasgos de este.

Measurement Phase
(1) For level $i \in \{\text{US}, \text{state}, \text{county}, \text{tract}, \text{block group}, \text{block}\}$ (a) Determine the privacy-loss budget for the level i; (b) Take differentially private noisy measurements $\widetilde{\mathbf{M}}_\gamma$ for all nodes in level i.
Estimation Phase
(1) For the US root node γ_0 estimate the contingency table vector $\mathbf{x}_{\gamma_0}$ by (a) Estimating a non-negative solution $\tilde{\mathbf{x}}_{\gamma_0}$ from the set of differentially private noisy measurements $\widetilde{\mathbf{M}}_{\gamma_0}$, invariants, and edit constraints at the US level; (b) Estimating a non-negative integer solution $\hat{\mathbf{x}}_{\gamma_0}$ from $\tilde{\mathbf{x}}_{\gamma_0}$ by controlled rounding. (2) For level $i \in \{\text{state}, \text{county}, \text{tract}, \text{block group}, \text{block}\}$, let P_i represent the set of distinct parents among all nodes at level i. For each parent node $\gamma \in P_i$, estimate the joint contingency table vector $\mathbf{x}_{\text{children}(\gamma)}$ by (a) Estimating a non-negative solution $\tilde{\mathbf{x}}_{\text{children}(\gamma)}$ from the set of differentially private noisy measurements $\widetilde{\mathbf{M}}_{\text{children}(\gamma)}$, invariants, edit constraints, and aggregate constraints enforcing that the children sum to $\hat{\mathbf{x}}_\gamma$; (b) Estimating a non-negative integer solution $\hat{\mathbf{x}}_{\text{children}(\gamma)}$ from $\tilde{\mathbf{x}}_{\text{children}(\gamma)}$ by controlled rounding.

Figura 5: Extracto del Algoritmo TopDown [1]

Se aprecia que el algoritmo se divide en 2 fases. La primera es la llamada "Measurement Phase" o fase de medición. En esta fase se itera por cada nivel del árbol geográfico, empezando por la raíz. En cada nivel se define un presupuesto de privacidad y se toman mediciones ruidosas para cada nodo del árbol en ese nivel, esto sin considerar ninguna restricción de valores negativos o que exista consistencia entre los valores generados entre un nivel y otro. Luego, en la "Estimation Phase", fase de estimación o también referida como fase de posprocesamiento, se utiliza lo encontrado en la fase anterior para encontrar una nueva solución. Esta debe parecerse a los valores ruidosos generados y satisfacer todos los invariantes y restricciones. Esto incluye, por ejemplo, que la población a nivel de estado sea la del valor real y que los nodos padres sean consistentes respecto a la agregación de sus nodos hijos.

Conceptos preliminares

Para la generación del ruido se utiliza una distribución Gaussiana discreta $\mathcal{N}_{\mathbb{Z}}(0, \sigma^2)$. En primer lugar, esto se debe a que se publican frecuencias, las cuales son valores enteros no negativos. En segundo lugar, a diferencia de una distribución de Laplace, la Gaussiana es menos "achatada" y tienes "colas" más chicas. Esto quiere decir que la probabilidad de que se generen valores más alejados a la media es menor. Esta decisión es tomada debido a que, para frecuencias de menor magnitud, se busca que sea poco probable que el ruido altere en gran medida el valor real, logrando mantener así mejor la utilidad. Por último, esta distribución satisface privacidad diferencial aproximada y también privacidad diferencial concentrada [2].

En lo que respecta al uso de la privacidad diferencial, el algoritmo TopDown emplea el mecanismo de "zero concentrated differential privacy (zCDP)". Este pertenece al grupo de mecanismos de privacidad diferencial aproximada, que, a diferencia de la privacidad diferencial tradicional, utiliza dos parámetros: un presupuesto de privacidad ε y una probabilidad de fallo δ. Este enfoque relaja las condiciones de privacidad al permitir un ε más pequeño, introduciendo a cambio una probabilidad δ de que no se cumpla la desigualdad de privacidad asociada y se rompa la privacidad diferencial [12].

La zCDP, en particular, conecta con la privacidad diferencial aproximada al utilizar un único parámetro ρ, el cual puede ser convertido en (ε, δ) a través de la siguiente ecuación:

$$\varepsilon = \rho + 2\sqrt{\rho \log(1/\delta)}$$

Este mecanismo simplifica el manejo del presupuesto de privacidad y satisface las garantías de privacidad establecidas por (ε, δ)-DP. Considerando esto, cuando se define un presupuesto de privacidad para cada nivel del árbol geográfico, se refiere a que se establece un ρ para cada nivel del árbol y no un ε.

Algoritmo

Ahora, tras aclarar estos conceptos se considerará el diagrama de la figura 6 para aterrizar los distintos conceptos del algoritmo. Este muestra un árbol que representa un ejemplo de jerarquización de la Universidad de Chile. Las tablas de contingencia o histogramas al costado de cada nodo muestran la frecuencia de hombres y mujeres. Los valores que están en azul representan un ejemplo del valor real. Se puede notar que la agregación de las tablas de contingencia de los nodos hijos suman a los del nodo padre para los valores reales. Por otro lado, en color rojo, se tiene el valor ruidoso que sería generado tras agregarle ruido en la fase de medición del algoritmo. Aquí se puede notar que la agregación de las mediciones ruidosas no son consistentes con las mediciones del nodo padre. Además, se puede ver que en el nodo de obstetricia, el valor ruidoso de la frecuencia de hombres quedó con un valor negativo de -1, esto es posible dada la aleatoriedad y la no imposición de restricciones en la fase de medición del algoritmo.

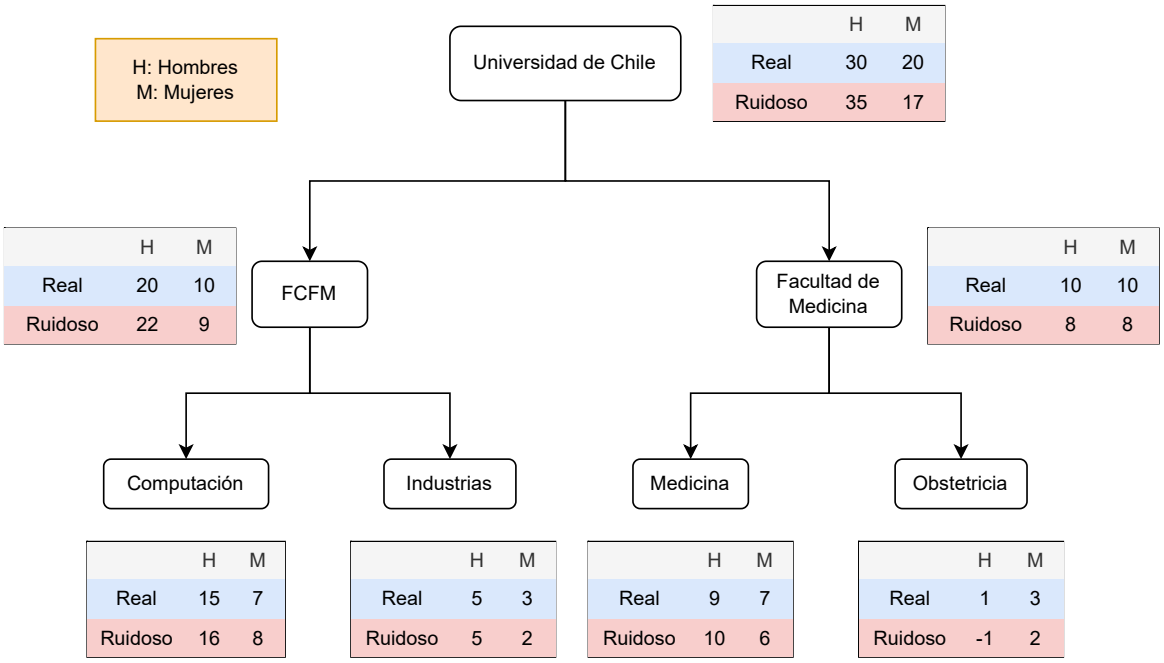

Figura 6: Árbol jerárquico de ejemplo con frecuencias reales y ruidosas

A continuación, utilizaremos y explicaremos la notación empleada en [1] para entender el algoritmo de la figura 5 con el ejemplo de la figura 6.

Primero, denotaremos el arreglo $\rho = [0.25, 0.5, 1]$ como el que contiene los presupuestos de privacidad asociados a cada nivel del árbol, empezando por la raíz. Se usa un presupuesto menor para la raíz ya que la magnitud de las frecuencias es mayor, en cambio, para nodos más cercanos a las hojas, cada vez es necesario tener mayor precisión a la hora de consultar los datos, lo que hace necesario utilizar un presupuesto mayor.

También se representa el árbol con el símbolo Γ, donde su raíz es γ_0 y los nodos del árbol $\gamma \in \Gamma$. Para referirse al conjunto de nodos hijos de un nodo γ se utilizará $children(\gamma)$. Por ejemplo, $children(\gamma_0) = \{\gamma_{FCFM}, \gamma_{FdM}\}$. A las hojas de un nodo se le llamarán $leaves(\gamma)$ y al padre de un nodo $parent(\gamma)$. Esta notación será relevante cuando se presenten ecuaciones.

También se comprimen los microdatos reales de una matriz $A \in \mathbb{Z}^{n \times k}$ a una tabla de contingencia o vector x. El largo de tal vector depende de qué atributos se consideren para su generación. En este caso podemos considerar todos los valores distintos de los datos, los cuales son:

- **Departamentos.** 4 posibilidades: las hojas del árbol.
- **Sexo.** 2 posibilidades: hombre o mujer.

Notemos que no se considera los 7 nodos del árbol debido a que se pueden ir agregando los hijos, partiendo de las hojas, para formar los valores internos del árbol. Puede pensarse

como el equivalente a que una columna entregue la comuna, y que a partir de esta se puedan agregar para obtener las regiones y el país. Por otro lado, en la práctica, la magnitud de todas las posibles permutaciones entre los valores distintos de todas columnas de los microdatos haría que el costo algorítmico se dispare considerablemente, por lo que se deberá escoger un subconjunto de atributos con los que trabajar.

Ahora, la cantidad de posibles combinaciones sería $c = 4 * 2 = 8$ debido a que hay 2 posibles sexos y 4 posibles geografías en las hojas del árbol. De manera similar, para un γ_i dado se tiene que $c^* = 2$, ya que al fijar γ_i no se considera la geografía para el cálculo de c.

Considerando estas definiciones, ahora el vector x que contiene los valores reales tiene un largo c. Donde el vector puede asumirse que esta ordenado de cierta manera para que describa en su totalidad a la tabla de contingencia. Por ejemplo, podría definirse el vector:

$$x = [15, 7, 5, 3, 9, 7, 1, 3]$$

Donde el orden se consigue al iterar por las hojas del árbol, de izquierda a derecha, y obteniendo las frecuencias de hombres y mujeres en cada nodo. Así el primer 15 corresponde a la cantidad de hombres en el Departamento de Ciencias de la Computación (DCC), el primer 7 a la cantidad de mujeres, el primer 5 a la cantidad de hombres del Departamento de Ingeniería Industrial (DII), y así sucesivamente. De igual forma, podemos denotar como x_γ como el vector para un nodo γ en específico. Por ejemplo, para $\gamma =$ FCFM, se tendría el vector $x_\gamma = [20, 10]$ de largo $c^* = 2$.

Junto a esto, se debe definir un conjunto de consultas lineales Q que se harán sobre x, tal que el producto matricial Qx entregue los resultados de tales consultas. La representación será $Q \in \{0, 1\}^{a \times c}$ para el árbol, o bien, $Q_\gamma \in \{0, 1\}^{a \times c^*}$ para un nodo γ fijo. Donde a representa la cantidad de consultas que se harán.

Considerando lo anterior, para el ejemplo podemos definir la matriz Q_γ como:

$$Q_\gamma = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Donde la primera fila representa la consulta por el total de personas, la segunda por el total de hombres y la tercera por el total de mujeres para un nodo γ en especifico.

Esto se puede generalizar para el árbol en su totalidad, se podría obtener las mismas consultas al concatenar las matrices Q_γ del nivel inferior de manera horizontal, o bien, se podría rescatar la misma consulta del nivel inferior al concatenar de manera vertical y rellenando con ceros de manera horizontal. Haciendo estas 2 operaciones se obtiene una matriz que representa todas las consultas del árbol. Por ejemplo, para el árbol Γ_{FCFM}, tendríamos:

$$Q_{\Gamma_{FCFM}} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Donde la primera fila es la consulta por el total de personas a nivel FCFM, la segunda la cantidad de hombres, la tercera la cantidad de mujeres, la cuarta la cantidad de personas del DCC, la quinta la cantidad de hombres en el DCC, y así sucesivamente.

Considerando esto, se puede representar el vector con el resultado de las consultas simplemente como Qx. Por ejemplo, para el nodo γ_{DCC}, tendriamos:

$$\begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 15 \\ 7 \end{bmatrix} = \begin{bmatrix} 22 \\ 15 \\ 7 \end{bmatrix}$$

Finalmente, teniendo en consideración toda esta notación, podemos definir el vector que agrega ruido a los resultados de los consultas como y. Donde $y \sim \mathcal{N}_{\mathbb{Z}}(0, \sigma^2)$ debido a lo explicado en la sección anterior.

Así el vector que contiene los valores ruidosos se puede denotar como $\widetilde{M} = Qx + y$. En particular, en el ejemplo; $\widetilde{M}_{DCC} = \begin{bmatrix} 24 \\ 16 \\ 8 \end{bmatrix}$.

Fase de Medición

Teniendo lo anterior en cuenta, la fase de medición es simple. Se itera por cada nivel del árbol geográfico, en este caso, a nivel universidad, luego a nivel de facultades y finalmente a nivel departamentos. En cada nivel se asigna un presupuesto de privacidad, que este caso es el arreglo $\rho = [0.25, 0.5, 1]$ que definimos anteriormente. Luego, se toman mediciones ruidosas para cada nodo de ese nivel, es decir, se toma la tabla de contingencia x_γ, se hacen las consultas $Q_\gamma x_\gamma$ y se le agrega el ruido y_i considerando el presupuesto ρ_i. Obteniendo de esta manera, en cada nodo del árbol, un vector $\widetilde{M}_\gamma = Q_\gamma x_\gamma + y_i$ que es el que esta representado en rojo en la figura 6.

Fase de Estimación

Una vez terminada la fase de medición, se empieza la fase de estimación o posprocesamiento. Esta itera de manera parecida a la fase anterior, pero en cada nivel lo que hace es estimar

una solución que se parezca "lo más posible" a cada $\widetilde{M}_\gamma$, pero que satisfaga invariantes de igualdad, desigualdad y sea consistente con los nodos del padre.

Los invariantes están relacionados a requisitos de publicación. Es en este punto donde se impone que la población estatal tiene que ser publicada con el valor real, esto debido a las políticas de transparencia del mandato constitucional [1]. También, en el caso del censo de viviendas, la cantidad de personas en un establecimiento debe ser mayor que 0. Además es aquí cuando se preocupa que el árbol sea consistente en sus valores, es decir, que la agregación de los nodos hijos sumen a los del padre considerando las soluciones encontradas.

Ahora, por temas de complejidad computacional, en cada nodo se divide la estimación en encontrar 2 estimaciones distintas. En primer lugar, se utiliza un estimador de mínimos cuadrados para encontrar una solución real no negativa que cumpla los invariantes, para luego, utilizar un estimador de redondeo para discretizar la solución.

Primero, el nodo raíz se procesa por separado, esto debido a que no tiene padre con el cual ser consistente. Entonces, se busca primero una solución real no negativa que cumpla los invariantes a través de una estimación de mínimo cuadrados. Se plantea el siguiente problema de optimización:

$$\begin{aligned} \tilde{x} \leftarrow \arg\min_{x_\gamma} & \left((Qx_\gamma - \widetilde{M}_\gamma)^\top W (Qx_\gamma - \widetilde{M}_\gamma) \right) \\ & \text{sujeto a:} \\ & x_\gamma \geq 0, \\ & C^{eq} x_\gamma = c^{eq}, \\ & C^{u} x_\gamma \leq c^{u} \end{aligned} \tag{1}$$

De este problema, la matriz W representa una matriz de pesos que contiene la varianza inversa de las variables aleatorias diferencialmente privadas a lo largo de su diagonal [1]. También, se da cuenta que lo que se busca encontrar es, entre todos los posibles vectores o tablas de contingencias x_γ no negativos que satisfacen las restricciones planteadas, aquel que tiene los resultados a las consultas Q más parecidos a las respuestas ruidosas $\widetilde{M}_\gamma$ calculadas en la fase de estimación.

Luego, la primera restricción es la que impone que todos los valores deben ser no negativos. La segunda restricción impone que se cumplan los invariantes de igualdad, como por ejemplo, que la población a nivel sea efectivamente la real. En este caso, podemos suponer que la cantidad de personas a nivel facultad debe ser transparente y publicar el valor real. Así, $c^{eq} = \begin{bmatrix} 30 \\ 20 \end{bmatrix}$ podría representar el vector con los valores reales, en este caso.

Sin embargo, a nivel de la raíz esto aun no tiene mucho sentido, ya que al estar fijando el nodo, la tabla de contingencia de la raíz x_{γ_0} no tiene información de las personas por facultad, sino solo a nivel de universidad. Equivalentemente en Estados Unidos, solo se tiene la información a nivel nacional. Por lo que tal valor de c^{eq} será útil cuando se vaya a un nivel del árbol más bajo. No obstante, en la práctica se pueden definir otras restricciones de

igualdad si los datos lo requieren.

Luego, C^{eq} sería parecida a Q, ya que es la matriz que transforma x_γ en c^{eq}, en otras palabras, consulta x_γ para transformarlo en el vector c^{eq}. La misma explicación aplica a la última restricción, la cual tiene que ver con las desigualdades. Un ejemplo de esto es que un establecimiento no puede tener 0 personas en Estados Unidos en el censo de viviendas. Así, C^u y c^u se construyen de manera similar a C^{eq} y c^{eq}.

En este punto del algoritmo en el cual se esta rompiendo la privacidad diferencial, ya que se están transformando los resultados ruidosos a unos datos más convenientes para la publicación. Es más, al imponer las restricciones de igualdad, se esta encontrando una solución que no utiliza privacidad diferencial para los datos expuestos en c^{eq}. Podría entenderse que para estos valores se le asigna un presupuesto de privacidad infinito [1], es decir, optando por la utilidad máxima a cambio de sacrificar la privacidad.

Por otro lado, podría argumentarse que, por las propiedades de posprocesamiento de la privacidad diferencial, esto no es cierto. Esto es una discusión que no se abordará actualmente, pero durante el próximo semestre podría ser interesante investigar e indagar al respecto.

Una vez resuelto este problema, se obtiene entonces un $\tilde{x}$ que representa una solución real no negativa que cumple con las restricciones, pero debido a que se publican frecuencias, ahora se necesita recurrir a un estimador de redondeo para discretizar la solución.

De manera similar para el cálculo de la solución de $\tilde{x}$, se plantea un problema de optimización para encontrar la solución no negativa entera que más se parezca a $\tilde{x}$:

$$
\begin{aligned}
\hat{x}_\gamma \leftarrow & \lfloor \tilde{x}_\gamma \rfloor + \hat{y}, \\
& \hat{y} = \arg\min_{y} \left(1^\top \left| \tilde{x}_\gamma - \left(\lfloor \tilde{x}_\gamma \rfloor + y \right) \right| \right) \\
& \text{sujeto a:} \\
& y_i \in \{0, 1\} \text{ para } y_i \text{ un elemento de } y; \\
& C^{eq}\left(\lfloor \tilde{x}_\gamma \rfloor + y \right) = c^{eq}; \\
& C^{u}\left(\lfloor \tilde{x}_\gamma \rfloor + y \right) \leq c^{u}.
\end{aligned}
\tag{2}
$$

De esto se puede notar que, $\hat{x}_\gamma$ sería la solución final para el nodo raíz, la cual cumple con todas las restricciones. Luego, la expresión $\lfloor \tilde{x}_\gamma \rfloor + \hat{y}$ hace referencia a que, de la solución encontrada en el paso anterior, se trunca todos los valores y se les suma un valor $y_i \in \{0, 1\}$, discretizando la solución. Se debe notar que el valor que es encontrado en esta estimación no es $\tilde{x}_\gamma$, sino que es el valor $\hat{y}$ que dice a que elementos de $\lfloor \tilde{x}_\gamma \rfloor$ se les suma 1 o no. Luego, las restricciones son las esperables, ya que son idénticas al estimador de mínimos cuadrados considerando que $\lfloor \tilde{x}_\gamma \rfloor + y$ representa la solución que se encontrará con esta segunda estimación.

Con esto completado, finalmente el algoritmo entra en un último paso, el cual es iterar por todos los niveles del árbol exceptuando la raíz. Donde, para cada nodo γ de un nivel del árbol dado, se resuelven los mismos problemas de estimación, solo que esta vez se agrega

una condición adicional; que los nodos sean consistentes con las soluciones calculadas por su padre. Para esto, es más sencillo pensar que se itera por los nodos padre $\gamma \in P_i$, donde P_i es el conjunto que representa a todos los nodos padres del nivel i del árbol, y se calcula la tabla de contingencia conjunta $x_{children(\gamma)}$.

Se puede entender $x_{children(\gamma)}$ como el resultado de apilar los valores de x_{child} para $child \in children(\gamma)$. Por ejemplo, considerando el árbol de ejemplo de la figura 6:

$$x_{children(FCFM)} = \begin{bmatrix} 15 \\ 7 \\ 5 \\ 3 \end{bmatrix}$$

Luego, de manera similar se puede definir $\widetilde{M}_{children(\gamma)}$ como el mismo vector que $x_{children(\gamma)}$, pero que contiene la agregación de los valores ruidosos de los nodos hijos en vez de los valores reales. Utilizando nuevamente la figura 6:

$$\widetilde{M}_{children(FCFM)} = \begin{bmatrix} 16 \\ 8 \\ 5 \\ 2 \end{bmatrix}$$

A continuación se plantean los estimadores modificados teniendo en cuenta la nueva restricción de consistencia:

$$\begin{aligned}
\tilde{x}_{children(\gamma)} \leftarrow & \arg\min_{x_{children(\gamma)}} \left(\left(Qx_{children(\gamma)} - \widetilde{M}_{children(\gamma)}\right)^\top W \left(Qx_{children(\gamma)} - \widetilde{M}_{children(\gamma)}\right) \right), \\
& \text{sujeto a:} \\
& x_{children(\gamma)} \geq 0; \\
& C^{eq} x_{children(\gamma)} = c^{eq}; \\
& C^{u} x_{children(\gamma)} \leq c^{u}; \\
& \begin{bmatrix} I_{c_*} & I_{c_*} & \cdots & I_{c_*} \end{bmatrix} x_{children(\gamma)} = \hat{x}_\gamma. \\
\hat{x}_{children(\gamma)} \leftarrow & \lfloor \tilde{x}_{children(\gamma)} \rfloor + \hat{y}, \\
& \hat{y} = \arg\min_{y} \left(1^\top \left| \tilde{x}_{children(\gamma)} - \left(\lfloor \tilde{x}_{children(\gamma)} \rfloor + y \right) \right| \right), \\
& \text{sujeto a:} \\
& y_i \in \{0, 1\} \text{ para } y_i \text{ un elemento de } y; \\
& C^{eq} \left(\lfloor \tilde{x}_{children(\gamma)} \rfloor + y \right) = c^{eq}; \\
& C^{u} \left(\lfloor \tilde{x}_{children(\gamma)} \rfloor + y \right) \leq c^{u}; \\
& \begin{bmatrix} I_{c_*} & I_{c_*} & \cdots & I_{c_*} \end{bmatrix} \left(\lfloor \tilde{x}_{children(\gamma)} \rfloor + y \right) = \hat{x}_\gamma.
\end{aligned} \tag{3}$$

No es difícil notar que las estimaciones son equivalentes a las anteriores, sin embargo, la última restricciones es la que hay que analizar con más cuidado.

Una restricción de consistencia podría ser, por ejemplo, que la cantidad de hombres de la FCFM coincida con la suma de los hombres en Computación e Industrias. Ahora, para establecer las restricciones de consistencia, se utiliza además de $\tilde{x}_{children(\gamma)}$ la matriz identidad I_{c_*}. Tal matriz tiene dimensión c^* debido a que el vector $\tilde{x}_{children(\gamma)}$ tiene dimensión $c^* \times |children(\gamma)|$. Si tomamos como ejemplo el nodo $\gamma = FCFM$ tendremos que $c^* = 2$. Considerando $x_{children(\gamma)}$ como lo expuesto con anterioridad, se concatenan tantas matrices I_{c_*} como nodos hijos tenga γ. Esto hace que la restricción tenga una matriz de concatenación de identidades de dimensión $c^* \times c^*|children(\gamma)|$ y por lo tanto, el nodo padre tenga dimensión $c^* \times 1$ que es justamente lo que se busca, ya que es la dimensión de su tabla de contingencia. Aplicando esto al ejemplo se tiene:

$$\begin{bmatrix} I_{c_*} & I_{c_*} \end{bmatrix} x_{children(\gamma)} = \hat{x}_\gamma$$

$$\begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 15 \\ 7 \\ 5 \\ 3 \end{bmatrix} = \begin{bmatrix} 20 \\ 10 \end{bmatrix}$$

Donde $\hat{x}_\gamma$ es exactamente la tabla de contingencia de la FCFM.

Una vez se termina de iterar por todo el árbol, significa que ya se tiene una solución $\hat{x}_\gamma$ para cada nodo γ. Esto es suficiente para dar por terminado el proceso de privatización y luego se pueden tomar todas estas soluciones para publicar los datos según se requiera.

Referencias

[1] Abowd, John M., Robert Ashmead, Ryan Cumings-Menon, Simson Garfinkel, Micah Heineck, Christine Heiss, Robert Johns, Daniel Kifer, Philip Leclerc, Ashwin Machanavajjhala, Brett Moran, William Sexton, Matthew Spence y Pavel Zhuravlev: *Experiments in Differential Privacy for the 2020 Census of Population and Housing*, 2022. https://www2.census.gov/adrm/CED/Papers/CY22/2022-002-AbowdAshmeadCumingMenonGarfinkelEtal.pdf, U.S. Census Bureau. Último acceso: 2 de diciembre de 2024.

[2] Canonne, Clément L, Gautam Kamath y Thomas Steinke: *The Discrete Gaussian for Differential Privacy*. En Larochelle, H., M. Ranzato, R. Hadsell, M.F. Balcan y H. Lin (editores): *Advances in Neural Information Processing Systems*, volumen 33, páginas 15676–15688. Curran Associates, Inc., 2020. https://proceedings.neurips.cc/paper_files/paper/2020/file/b53b3a3d6ab90ce0268229151c9bde11-Paper.pdf.

[3] Dalenius, Tore y Stephen P. Reiss: *Data-Swapping: A Technique for Disclosure Control*. Journal of Statistical Planning and Inference, 6:73–85, 1982.

[4] Dwork, Cynthia y Aaron Roth: *The Algorithmic Foundations of Differential Privacy*. Found. Trends Theor. Comput. Sci., 9(3-4):211–407, 2014. https://doi.org/10.1561/0400000042.

[5] European Commission: *Handbook on Statistical Disclosure Control*, 2023. https://cros.ec.europa.eu/system/files/2023-12/SDC_Handbook.pdf, Último acceso: 28 de noviembre de 2024.

[6] Fraser, Bruce y Janice Wooton: *A Proposed Method for Confidentialising Tabular Output to Protect Against Differencing*, 2005. https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2005/wp.35.e.pdf, Supporting Paper, submitted by the Australian Bureau of Statistics. Joint UNECE/Eurostat work session on statistical data confidentiality, Geneva, Switzerland, 9-11 November 2005.

[7] Fung, Benjamin C.M., Ke Wang, Ada Wai Chee Fu y Philip S. Yu: *Introduction to Privacy-Preserving Data Publishing: Concepts and Techniques*. Chapman & Hall/CRC, 1st edición, 2010, ISBN 1420091484.

[8] Garfinkel, Simson: *Differential Privacy and the 2020 US Census*. MIT Case Studies in Social and Ethical Responsibilities of Computing, (Winter 2022), jan 24 2022. https://mit-serc.pubpub.org/pub/differential-privacy-2020-us-census.

[9] Instituto Nacional de Estadísticas, Chile: *Manual para Censistas - Censo 2017*, 2017. https://www.ine.gob.cl/docs/default-source/censo-de-poblacion-y-vivienda/manuales/censo-2017/manual-para-censistas.pdf?sfvrsn=80f7e0c9_4, Último acceso: 4 de septiembre de 2024.

[10] Instituto Nacional de Estadísticas (INE): *Censo de Población y Vivienda 2017 - Microdatos*, 2017. https://www.ine.gob.cl/estadisticas/sociales/

censos-de-poblacion-y-vivienda/censo-de-poblacion-y-vivienda, Contiene información sobre los microdatos del Censo de 2017 en Chile. Último acceso: 4 de septiembre de 2024.

[11] Instituto Nacional de Estadísticas (INE): *Manual de Usuario de la Base de Datos del Censo de Población y Vivienda 2017*. Instituto Nacional de Estadísticas (INE), 2018. https://www.ine.gob.cl/docs/default-source/censo-de-poblacion-y-vivienda/bbdd/censo-2017/manual-de-usuario/manual_de_usuario_censo_2017_16r.pdf?sfvrsn=38710602_4, Último acceso: 1 de diciembre de 2024.

[12] Near, Joseph P. y Chiké Abuah: *Programming Differential Privacy*, volumen 1. 2021. https://programming-dp.com/.

[13] República de Chile: *Ley N° 19.628 sobre Protección de la Vida Privada*, 1999. https://www.bcn.cl/leychile/navegar?idNorma=141599, Promulgada el 28 de agosto de 1999.

[14] República de Chile: *Ley N° 20.285 sobre Acceso a la Información Pública*, 2008. https://www.bcn.cl/leychile/navegar?idNorma=276363, Promulgada el 11 de agosto de 2008.

[15] República de Chile: *LEY NÚM. 20.840 SUSTITUYE EL SISTEMA ELECTORAL BINOMINAL POR UNO DE CARÁCTER PROPORCIONAL INCLUSIVO Y FORTALECE LA REPRESENTATIVIDAD DEL CONGRESO NACIONAL*, 2015. https://www.bcn.cl/leychile/navegar?idNorma=1077039, Promulgada el 27 de abril de 2015.

[16] Statistics Netherlands: *Statistical Disclosure Control*, 2023. https://research.cbs.nl/casc/index.htm, Online resource for statistical confidentiality tools and research.

[17] Toro, Matías: *Técnicas formales de privacidad de datos: ¿Está el Servel protegiendo nuestra privacidad?* Revista Bits de Ciencia, 1(22):43–48, 2022. https://revistasdex.uchile.cl/index.php/bits/article/view/12645/12666.

[18] United States Congress: *U.S. Code, Title 13 - Census*, 2009. https://www.govinfo.gov/content/pkg/USCODE-2009-title13/html/USCODE-2009-title13.htm.

[19] U.S. Census Bureau: *About Congressional Apportionment*, 2021. https://www.census.gov/topics/public-sector/congressional-apportionment/about.html, Último acceso: Noviembre 25, 2024.