LUCHANDO CONTRA EL COVID19 CON DATOS

14 de marzo 2020

Autores (a título personal desvinculado de nuestra labor profesional)
Daniel Villatoro, Nuria Oliver, Elena Alfaro, Alejandro Llorente, Juan Murillo, Giovanna Miritello, Irene Rodríguez-Luján, Oriol Pujol, Curro Maturana.
(+ quien sea que colabore por favor que añada su nombre en un comentario)

Objetivo. Presentar una solución técnica para la identificación del grado de exposición personal al COVID19 mediante una aproximación basada en dato empírico de operadoras telefónicas. Esta propuesta apunta a la creación de un grupo de trabajo centralizado y gubernamental que trabaje con los datos reales de operadoras telefónicas y datos de Sanidad para calcular zonas de riesgo de contagio, una probabilidad personalizada del grado de exposición en base a los positivos que se vayan detectando, y todo asegurando la privacidad de los individuos y los datos de cada una de las entidades colaboradoras.
Esta aproximación, contando con el apoyo de todas las operadoras telefónicas que operan en España, resultaría más completa (al abarcar a toda la población con móvil) que el resto de aproximaciones colaborativas que están siendo exitosas en otros países con más grado de digitalización (como Corona 100m en Corea).

PLANTEAMIENTO DEL PROBLEMA Y OBJETIVO

El COVID19 se transmite a través del contacto personal, y puede resultar asintomático hasta transcurridos 15 días. Puede haber también transmisión a través de objetos que hayan sido infectados previamente.

Debido a la invisibilidad del proceso de infección, es difícil identificar las personas que han estado en contacto con los infectados durante períodos asintomáticos de infección del virus, pero no imposible y más adelante presentamos una propuesta para resolver este problema.
Dicha característica está provocando la expansión del virus de manera silenciosa entre portadores asintomáticos, y la paranoia en ciudadanos aprensivos con posibilidades reducidas de estar infectados por falta de exposición a personas infectadas.

Por ejemplo, Ana estuvo todo el lunes con Berto y Carlos, cuando todos aparentaban estar sanos. Sin embargo, Ana empieza a encontrarse mal el martes, mientras que Berto y Carlos siguen OK. Es fundamental que Ana pueda recibir atención médica y ser testeada sobre el virus. En caso de resultar positiva, es fundamental que Berto y Carlos pasen a observación debido a su interacción directa con Ana durante un período de difusión de la enfermedad, y a reducir sus interacciones con otras personas para rebajar la posibilidad de continuar distribuyendo el virus. En caso que Ana no resulte positiva, Berto y Carlos pueden mantener la calma.

Mientras tanto, Davinia sigue encerrada en casa desde hace 5 días pese a no haber estado en contacto directo con ningún infectado en días anteriores a su reclusión. Esta reclusión pese a que es positiva respecto a su prevención, podría no ser tan estricta en caso de necesitar ayudar a terceros.

Por lo tanto para poder controlar la expansión del virus, es fundamental identificar de manera rápida y adecuada (1) personas infectadas de COVID19, (2) las personas que han estado en contacto con ellos durante el período de infección, (3) y los lugares donde han estado los infectados para poder higienizarse adecuadamente.

Adicionalmente, en el escenario actual, un análisis profundo de este tipo de red y datos geolocalizados también puede ayudar a objetivos como:

1. Cuantificación del cumplimiento real de medidas de confinamiento domiciliario
2. Cuantificación del impacto de las medidas en la propagación de la pandemia
3. Detección de focos de incumplimiento (ejemplo: detección de empresas que podrían teletrabajar y están obligando a sus empleados a ir a las oficinas)
4. Impacto de la movilidad remanente (viajes de trabajo, supermercados, estancos)
5. Medición de la evolución intra-municipios (a día de hoy, estudiando las curvas de distintas regiones, deja de ser útil la curva agregada en España; muchas empiezan a tener un número significativo de casos con posibles diferencias en la velocidad de propagación).

Identificar a las personas que han estado en contacto con infectados por el virus tiene un beneficio directo sobre la adecuada gestión y priorización de los servicios médicos y de tests, con el objetivo de “aplanar la curva” en los servicios sanitarios.

Informar a las personas que han estado en contacto con los portadores del virus durante el período de infección invisible resulta de ayuda para gestionar de manera inteligente la asistencia médica; adicionalmente, informar a aquellas personas que no han sido expuestas al virus permite transmitir un mensaje de calma. Finalmente, tomar medidas específicas en los lugares que han visto agregación de muchos contactos portadores del virus (desinfección, cierre, etc) también permitirá adoptar medidas preventivas si es necesario.
Inspirado en la aplicación coreana Corona 100m, es importante también identificar los lugares y momentos que han tenido presencia del virus, para así minimizar el riesgo de exposición a esos lugares a las personas que los transitan habitualmente.

Estas medidas han sido efectivas para contener la propagación del Covid-19 en Corea y en Singapur. Dado el crecimiento exponencial del número de infectados en el territorio estatal, consideramos de suma importancia adoptar medidas de este tipo para evitar el colapso del sistema sanitario.

PROPUESTA DEL ANÁLISIS DE INFECCIONES MEDIANTE EL GRAFO DE INTERACCIONES

Mediante teoría de grafos, resulta trivial computar el grado de exposición de cada persona a otras personas o zonas contagiadas, y por lo tanto se podría calcular la probabilidad individual de infección. Esto es una área de conocimiento científico muy madura conocida como Social Contagion Processes, liderada por científicos como Alessandro Vespignani, Alex Sandy Pentland, Esteban Moro o Nuria Oliver.

Sin embargo, se presentan 3 retos importantes en este cometido: (1) la recopilación de datos de personas infectadas, (2) la construcción del grafo temporal de relaciones personales y localizaciones geográficas de los infectados, y (3) el adecuado proceso informativo a cada una de las personas en el momento preciso.

Es imposible capturar la realidad fehaciente del proceso completo debido a la imposibilidad de tener datos empíricos completos, sin embargo, existen posibilidades que pueden cubrir el problema de una manera masiva (pese a ser incompleta).

Pocas entidades poseen datos que nos permitan recopilar información sobre este proceso. Al mismo tiempo, los datos aportan un valor inmenso ya que sin datos es imposible hacer este tipo de análisis y predicciones.

Recientemente han surgido distintos movimientos civiles para compartir datos con todos los casos en tiempo real. Un ejemplo se puede ver aqui y aqui.

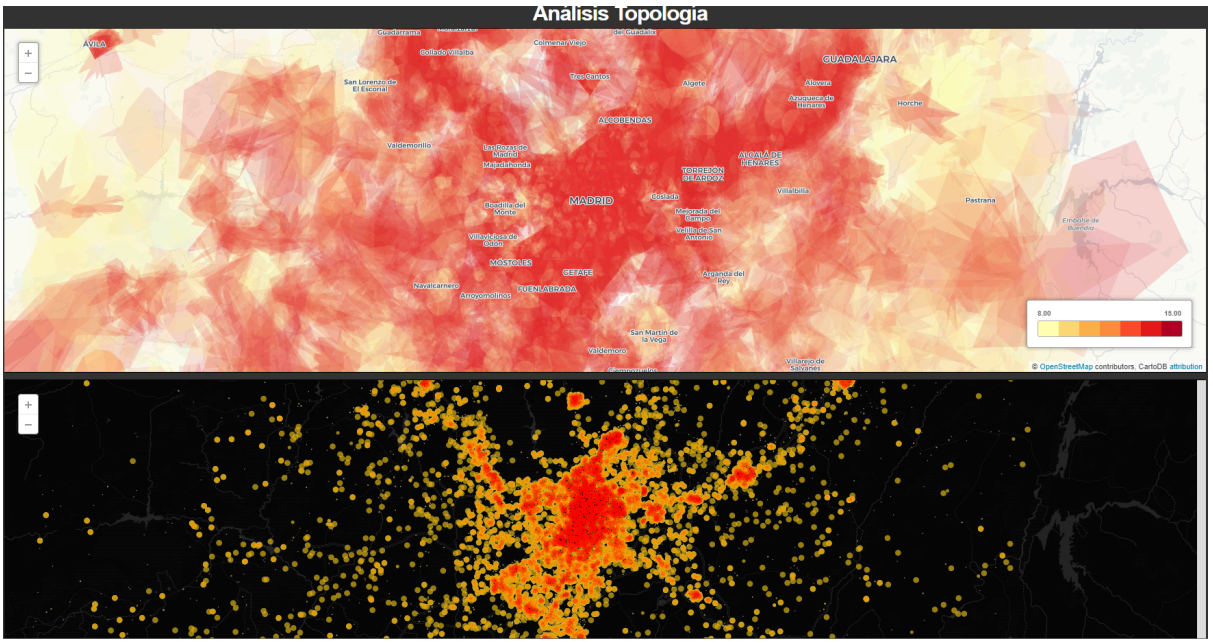
Además, no cabe duda de que una combinación de esfuerzos realizado por la Administración Pública y las Operadoras Telefónicas podría ayudar a atacar el problema de una manera analítica, basada en datos y en tiempo real, de la siguiente manera.[1]

Las operadoras telefónicas --así como Google y Facebook-- poseen dos de los 3 elementos fundamentales para abordar este reto: los datos que permiten construir el grafo temporal de movimientos en el mundo físico y la capacidad de comunicarse con los ciudadanos.

Para poder prestar el servicio de telefonía móvil, las operadoras telefónicas poseen una infraestructura de antenas fijas en el espacio que proveen de cobertura a nuestros dispositivos móviles. Estas antenas llegan a estar distribuídas en el espacio para poder proveer de cobertura de manera adecuada, y en las zonas de alta densidad llegan a estar a una distancia de 500m entre ellas.

Este es un ejemplo de la distribución de antenas 2G y 3G en Madrid (datos de hace algunos años)

[1] Una combinación de datos de Administraciones Públicas de Salud combinada con las grandes empresas de Internet con masa crítica de usuarios (como Google y Facebook) también sería de crucial utilidad aunque entendemos que a nivel gubernamental es más complicado gestionar el acceso a sus datos. Sin embargo es una posibilidad de la cual no debemos perder el foco ya que acometería la solución de la identificación a una escala global.



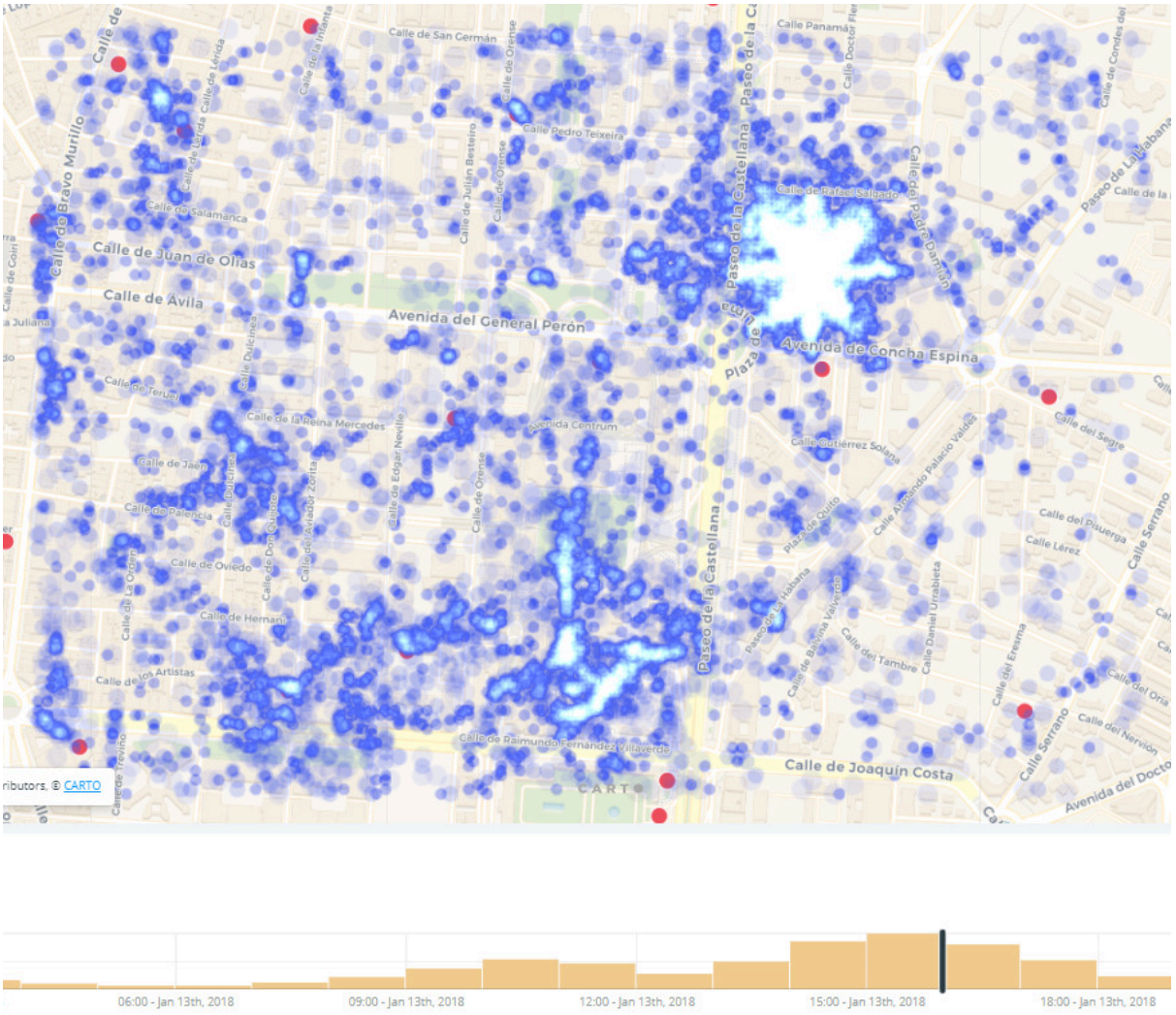
Por lo tanto, cada operadora guarda información sobre qué dispositivo está siendo provisto de cobertura por cada antena, en cada momento, y por lo tanto, conocer la localización aproximada de las portadores de dispositivos móviles en cada momento.

Asumiendo que cada persona tiene un dispositivo móvil, este factor posiciona a las operadoras (o a empresas como Google / Facebook) en la situación de conocer qué persona ha estado dónde y con quien en el mismo momento, con una precisión aproximada [2]. En el caso de las empresas de telecomunicaciones, esta información sólo es captada si el móvil está siendo utilizado o si la operadora tiene un 'network probe' que capte información incluso cuando los móviles no están siendo usados activamente. Gracias a las aplicaciones móviles con alta frecuencia de acceso a datos móviles (estilo Whatsapp) se tiene una granularidad detallada sobre la actividad diaria de los ciudadanos.

A continuación podemos ver dos ejemplos con distintos grados de granularidad del dato. En este gif están representados de los eventos de red donde se puede ver la actividad que llega a perfilar el AVE desde Barcelona hasta Zaragoza con bastante claridad.

[2] El datos de las celdas de cobertura de los operadores de telefonía pueden llegar a obtener una precisión geográfica del posicionamiento de un dispositivo con un error de 100m en las zonas de mayor densidad de celdas. En caso de querer obtener una precisión mayor será necesario abandonar este plan y optar por la solución estilo aplicación móvil que obtiene un posicionamiento GPS de mayor precisión.

En el siguiente ejemplo, vemos la zona del Bernabeu un día de partido:



Estos datos permitirían construir lo que se conoce como un grafo de co-location temporal, una estructura de datos que permite calcular quién ha estado con quién, en qué momento, y en qué lugar. Con esta estructura de datos, resultaría sencillo, una vez identificado un nuevo caso positivo de infectado por el virus, identificar qué otras personas y lugares han estado expuestos los días previos al test en la fase de infección invisible.

Información agregada de movilidad inferida a partir de este tipo de datos se hizo relevante recientemente por este estudio promovido por el INE.

Dicho estudio se realizó de manera coordinada por el INE con datos agregados de las 3 principales operadoras del país asegurando que no se creaban intereses comerciales competitivos, y asegurando la anonimización de las personas incluidas en dichos análisis.
En este caso, y para que esta iniciativa resulte exitosa es necesario contar con los datos de todas las operadoras del país[3] e idealmente también de Google/Facebook

[3] Los principales operadores de red (OMR) con cobertura propia en España son:

Por lo tanto, combinando los datos de todas las operadoras del país, nos permitirían identificar qué personas han estado en el mismo momento aproximadamente en el mismo entorno físico.

Sin embargo, es necesario que las autoridades sanitarias puedan informar qué personas están resultando positivas en COVID19, y así reconstruir esta traza de interacción de personas con infectadas, combinando el dato de las telcos y las autoridades sanitarias por primera vez.

En este punto es donde resulta imprescindible que los datos sean tratados de manera anónima pero con la capacidad de re-identificar a las personas con mayor grado de exposición y así poder informarles adecuadamente, y más adelante proponemos una solución técnica a la anonimización.

Al igual que en el informe del INE, es necesario que se establezca una iniciativa coordinada y ajena a las telco, donde las telco y las autoridades sanitarias provean sus datos de manera estructurada y organizada para poder realizar estos análisis, sin poner en riesgo la actividad comercial entre ellas o la privacidad de cada una de los ciudadanos.

Otra alternativa para las semanas venideras es el desarrollo e instalación masiva de una aplicación en el móvil similar a la aplicación coreana que permitiese a los ciudadanos reportar síntomas, confirmar diagnósticos así como acceder a sus localizaciones y contactos en los días previos a ser diagnosticados. Esta aproximación pese a ser más precisa en cuanto a la localización, tiene la desventaja de solo cubrir a los usuarios de dicha aplicación; mientras que con la aproximación telco se cubriría a toda la población con dispositivos móviles.

Esta iniciativa deberá ser responsable de:

- la construcción del grafo completo de interacción de los españoles en los últimas semanas, para así poder identificar de manera retrospectiva cada una de las personas expuestas al virus.
- la actualización de casos positivos sobre el grafo, para poder computar la probabilidad de infección de las personas que hayan estado expuestas al virus.
- la comunicación coordinada e individualizada a los ciudadanos con su grado de probabilidad de infección, para que así asistan a los centros médicos de manera sistematizada.

---

1. Grupo Movistar. Principales marcas: Movistar, Tuenti y O2
2. Grupo Orange. Principales marcas: Orange, Jazztel, Simyo y Amena
3. Grupo Vodafone. Principales marcas: Vodafone y Lowi
4. Grupo MásMóvil. Principales marcas: Yoigo, MásMóvil, PepePhone, Lebara y Llamaya
5. Grupo Euskaltel. Principales marcas: Euskaltel, R y Telecable

- la comunicación sobre las zonas geográficas con mayor grado de exposición y riesgo.

Para poder llevar a cabo esta iniciativa será necesario reunir un grupo de Data Scientists + Data Engineers junto a los equipos de Datos de las telcos y un responsable de los sistemas sanitarios del Gobierno Central y de las Comunidades Autónomas.

En esta iniciativa, estará compuestas de las siguientes tareas:

- Aprovisionamiento de los datos telco[4], histórico y diario para poder mantener actualizaciones.
- Aprovisionamiento de los datos de infectados, con los metadatos asociados como la fecha del test y la fecha de positivo de la infección.
- Cálculo del grafo de co-location.
- Cálculo de las matrices de movilidad.
- Cómputo de las zonas geográficas más expuestas a partir de la movilidad de los infectados.
- Cálculo de la probabilidad de infección a partir de nivel de exposición retrospectiva a infectados invisibles.
- Análisis descriptivo sobre la velocidad de infección, maduración del virus y propagación.
- Análisis descriptivo sobre la distribución demográfica de los casos teniendo en cuenta datos de censos y distribuciones en otros Países afectados por el virus, muy importante para entender cómo el virus se ha propagado y hasta qué punto las medidas tomadas habían podido ser más eficaces. Un análisis de los casos de Corea del Sur, Italia y otros Países de Europa se puede ver aquí además de interesantes simulaciones aquí.

Para poder acometer este proyecto es imprescindible:

- Acceso a los datos de las telco de manera estructurada, apalancando esfuerzos en sus propios equipos de data.
- Acceso a los datos de infectados geolocalizados y en una escala temporal.
- Infraestructura cloud para poder almacenar todos los datos.
- Infraestructura cloud para poder ejecutar todos los cómputos, mediante el uso de herramienta opensource como PySpark.
- Equipo de Data Engineers + Data Scientists + Developers + epidemiólogos / científicos

ANONIMIZACIÓN DE LOS DATOS

Llegados a este punto, la privacidad de cada uno de los ciudadanos resulta el punto más sensible a mantener y a asegurar, tanto para los ciudadanos como para cada uno de los operadores.

[4] E idealmente de Google y Facebook.

Según la ley actual de protección de datos que nos ampara (GDPR)[5], sólo bajo el consentimiento expreso de los ciudadanos se puede llegar a usar este dato de manera identificable. El éxito de esta iniciativa se fundamenta en el acceso completo al historial de interacciones de las ciudadanos y los lugares visitados; tomando una aproximación colaborativa donde los ciudadanos tengan que proveer sus datos voluntariamente (y por lo tanto aceptando el consentimiento sobre la cesión y uso de sus datos) reduce drásticamente el potencial éxito de esta iniciativa. Construyendo sobre el ejemplo anterior de los amigos Ana, Berto y Carlos: si Ana da positivo (y además ha cedido sus datos), solo si Berto y Carlos han aceptado el consentimiento podrán llegar a ser contactados como personas expuestas al virus y con cierto grado de riesgo de infección.

En el proceso de anonimización de datos, tenemos 3 tipos de actores principales dentro de esta iniciativa: proveedores de datos de telco, proveedores de datos sanitarios, gestor gubernamental de la aplicación centralizada.

Es fundamental que el gestor gubernamental de la aplicación centralizada, que va a tratar la totalidad de los datos y gestionar la aplicación, no tengan la capacidad de reidentificación personal de los ciudadanos sin su consentimiento explícito.

Por lo tanto es fundamental que los proveedores de datos gestionen un sistema de anonimización de datos que sea robusto (asegurando la imposibilidad de reidentificación del individuo) y que sea consistente (manteniendo las individualidad de clientes entre entidades y que no haya colisión de identidades anónimas).

Por lo tanto, para asegurar el tratamiento de los datos de manera anónima se plantea el siguiente procedimiento que afecta a dos etapas: la primera de procesado del dato y construcción del grafo de viralidad, y la segunda de gestión de la aplicación ciudadana en tiempo real.

Anonimización para la Construcción del Grafo.

- Los proveedores de datos (telco y sanidad) asignaran a una persona como responsable del proceso de anonimización central y coordinado entre todas las entidades que ceden datos. Esta persona será la encargada de conocer el algoritmo de anonimización que se aplica a los identificadores de los ciudadanos, y que en este caso ha de ser forzosamente el DNI / NIE. Nuestra propuesta es aplicar un SHA512 con una semilla que solo sea conocida por el responsable del proceso de anonimización central y cada uno de los proveedores de datos. Es necesario que entre los proveedores de datos el

[5] (Actualización 15 de marzo 2020) La Agencia Española de Protección de Datos ha publicado una nota en la que GDPR permite *"el tratamiento de datos personales de salud sin consentimiento del interesado en situaciones de interés público en el ámbito de la salud pública y en el cumplimiento de obligaciones legales en el ámbito laboral derivado de dichas situaciones"*; esta nota ha sido publicada dentro del contexto de la crisis epidémica del COVID19. Sin embargo, creemos que el mecanismo propuesto en este documento es asequible de implementar y asegura los intereses comerciales de todas las entidades que cedan datos.

algoritmo y la semilla sea el mismo, y en ningún caso conocido por la contraparte gubernamental que analizará los datos. La información de la semilla del algoritmo de anonimización deberá ser marcada como Confidencial y deberá ser conocida por el menor número de personas.
- De esta manera, el dato ya queda pseudo-anonimizado y puede ser tratado por el equipo gubernamental de data scientists, sin tener la posibilidad de revertir la identidad de cada uno de los ciudadanos presentes en los datos y por lo tanto en el grafo.
- Es relevante resaltar que este grafo permitirá (de manera anónima y dato agregado) identificar (1) que zonas geográficas están más expuestas al virus y por lo tanto necesitan medidas más duras de contención, y (2) gracias a los patrones de movilidad anónimos identificar que zonas van a ser las siguientes más afectadas.

Gestión de la Anonimización en fase de Aplicación

- Con los datos anonimizados individuales que gestione el equipo gubernamental, ya es posible computar para cada nodo del grafo (y por lo tanto para cada ciudadano) cuál es la probabilidad empírica de exposición al virus de manera actualizada, mediante el uso de los datos actualizados de los equipos sanitarios. Sin embargo, en este punto el dato es aún anónimo e imposible de informar a los ciudadanos.
- Para que los ciudadanos puedan conocer su grado de exposición real al virus, es necesario re-identificarles en el grafo y por lo tanto obtener su consentimiento explícito. En este momento es donde gracias a un punto centralizado de comunicación (p.ej. una app gubernamental que se encargue de comunicar a los ciudadanos sobre el protocolo de actuación actual o la evolución de la epidemia) se debe recopilar este consentimiento.
- Una vez recopilado el consentimiento, la app gubernamental cederá a cada operador este consentimiento individual.
- Mediante un servicio de APIs publicado por el equipo gubernamental, que computa el riesgo de infección para cada nodo del grafo a partir de su identificador anonimizado, cada operador podrá anonimizar el identificador de cada cliente usando la mismo procedimiento que en el primera fase y consultar esa API.
- El operador devolverá al cliente mediante la app o un sistema alternativo más sencillo cual es la probabilidad computada de infección.

Este protocolo asegura que se puede computar el grafo completo de manera anónima y no reversible, en entornos estancos de acceso a datos: ningún operador tendrá acceso al dataset completo gestionado por el equipo gubernamental, y el equipo gubernamental nunca podrá revertir la información personal identificable de cada ciudadano.
Solo el expreso consentimiento del ciudadano obtenido a través de la app centralizada gubernamental provocará el procedimiento del cálculo de probabilidad, que se realiza siempre de manera anónima sin exponer ningún dato personal de terceros.