




Interaction data are identifiable even across long periods of time

Ana-Maria Crețu ^{1,2}, Federico Monti^{3,4}, Stefano Marrone^{1,2,5}, Xiaowen Dong⁶, Michael Bronstein^{1,3,4} & Yves-Alexandre de Montjoye ^{1,2} 

Fine-grained records of people's interactions, both offline and online, are collected at large scale. These data contain sensitive information about whom we meet, talk to, and when. We demonstrate here how people's interaction behavior is stable over long periods of time and can be used to identify individuals in anonymous datasets. Our attack learns the profile of an individual using geometric deep learning and triplet loss optimization. In a mobile phone metadata dataset of more than 40k people, it correctly identifies 52% of individuals based on their 2-hop interaction graph. We further show that the profiles learned by our method are stable over time and that 24% of people are still identifiable after 20 weeks. Our results suggest that people with well-balanced interaction graphs are more identifiable. Applying our attack to Bluetooth close-proximity networks, we show that even 1-hop interaction graphs are enough to identify people more than 26% of the time. Our results provide strong evidence that disconnected and even re-pseudonymized interaction data can be linked together making them personal data under the European Union's General Data Protection Regulation.

¹Department of Computing, Imperial College London, London SW7 2AZ, UK. ²Data Science Institute, Imperial College London, London SW7 2AZ, UK. ³Faculty of Informatics, Università della Svizzera Italiana, 6904 Lugano, Switzerland. ⁴Twitter, London W1B 5DL, UK. ⁵University of Naples Federico II, 80125 Naples, Italy. ⁶Department of Engineering Science, University of Oxford, Oxford OX2 6ED, UK. ✉email: demontjoye@imperial.ac.uk

An increasing fraction of our online and offline interactions are now captured by technology¹. Large amounts of interaction data are now collected by messaging apps, mobile phone carriers, social media companies, and other apps to operate their service or for research purposes. Interaction data typically consist of the pseudonyms of the interaction parties, the timestamp of the interaction, and possibly further information. Mobile phone interaction data have been used to study the linguistic divide in a country², to study the interaction patterns of individuals with close connections over time³, or to forecast the spatial spread of epidemics⁴. Similarly, interaction data have been used to study the spread of misinformation on Twitter^{5,6}, the characteristics of news retweet networks during elections⁷, or the effect of Facebook friendship ties in political mobilization⁸. Finally, close-proximity interaction data have been collected using Bluetooth to study human behavior^{9–11} and are currently at the core of COVID-19 contact tracing apps aiming to help control the spread of the disease.

Despite previous claims^{12,13}, interaction data are deeply personal and sensitive. They record with high precision who we talk to or meet, at what time, and for how long. Sensitive information can furthermore often be inferred from interaction data. Previous research, for instance, showed how algorithms can predict who a person's significant other is¹⁴, their wealth^{15,16}, demographics^{17,18}, the propensity to overspend¹⁹, personality traits²⁰, and other attributes²¹ from interaction data. Some works even leveraged homophily or network ties when making predictions²². Legal scholars and privacy advocates have long argued that interaction data are as sensitive as the content of the communication and that “metadata are data”^{23,24}. Mobile phone metadata have been at the core of the Snowden revelations and their collection was later deemed illegal in *ACLU vs. Clapper*^{25,26}. More recently, the proportionality of contact tracing apps developed in the context of the COVID-19 pandemic has been questioned^{27–29}.

Interaction data can be shared or sold to third parties without users' consent, so long as they are anonymized. According to current data protection regulations such as the European Union's General Data Protection Regulation (GDPR)³⁰, or the California Consumer Privacy Act (CCPA), anonymized (or de-identified) data are no longer considered as personal data. The European Data Protection Board (EDPB) predecessor, the Article 29 Working Party, defined anonymization as resistance to singling out, linkability, and inference attacks³¹. In particular, the linkability criterion refers to “the ability to link, at least, two records concerning the same data subject.” While guidances are subject to the interpretation of the courts, matching identities between two pseudonymous datasets would likely mean that they are not anonymous under GDPR. Both legislations emphasize that personal data should not be stored for longer than necessary and then deleted or anonymized, with terms of service suggesting the latter to be common practice^{32–34}.

Matching attacks have long been used to identify individuals in datasets using matching auxiliary information, calling into question their anonymity. In one seminal study, zip code, birth date, and gender were used to identify the Governor of Massachusetts William Weld³⁵; in another, the movies people had watched were used³⁶. In 2013, it was shown that four points, approximate places and times, were enough to uniquely identify someone in location data 95% of the time³⁷, with formal similarity measures being proposed for approximate matching³⁸. Numerous matching attacks have been proposed for interaction and graph data, both using exact^{39–46} or approximate^{47–54} matching information. Graph matching^{55–58} and anchor links prediction^{59,60} are two closely related problems.

We here propose a profiling attack for interaction data based on geometric deep learning⁶¹. While matching attacks rely on auxiliary information fairly stable over time (gender, zip code, etc.) or from the same time period (spatio-temporal points, movies watched, etc.), profiling attacks use auxiliary information from one time period to profile and identify a person in another non-overlapping time period. This makes them more broadly applicable, as the auxiliary data does not have to come from the same time period as the dataset.

Using a graph attention neural network⁶², we learn an individual's behavioral profile by building a vector representation (embedding) of their *weekly* k -hop interaction network. Our weekly profiles use only behavioral features, aggregating both node features and topological information typically present in interaction data, and are optimized for identification. In a mobile phone dataset of more than 40k people, our model was able to correctly identify a person 52% of the time based on their 2-hop interaction network ($k=2$). Using only a person's interactions with their direct contacts ($k=1$), our model could still identify them 15% of the time. We further show that the accuracy of our model only decreases slowly as time passes with 24% of the people still being correctly identified after 20 weeks ($k=2$), thus making identification a real risk in practice. Finally, we show that our general graph profiling approach can be applied to other types of interaction data. We apply our model to Bluetooth close-proximity data similar to the one collected by COVID-19 contact tracing apps for more than 500 people and show that it is able to link together 1-hop interaction networks with 26% accuracy. Our results provide evidence that disconnected and even re-pseudonymized interaction data remain identifiable even across long periods of time. These results strongly suggest that current practices may not satisfy the anonymization standard set forth by the EDPB in particular with regard to the linkability criteria.

Results

Setup. Our attack exploits the stability over time of people's interaction patterns to identify individuals in a dataset of interactions using auxiliary k -hop interaction data from a disjoint time period.

We consider a service S collecting data about the interactions it is mediating. We denote by \mathcal{I} the set of individuals taking part in the communications recorded by S . For example, \mathcal{I} could be the set of users of a contact tracing or messaging application or the subscribers of a mobile phone carrier and their contacts. We call interaction data the record describing the interaction between two individuals using S , consisting of the pseudonym of the two individuals, a timestamp, and sometimes other information. We define a time period $\mathcal{T} = [t, t')$ as the set of all timestamps between a start t (inclusive) and end t' (exclusive). Given a time period \mathcal{T} , we define the interaction graph $G_{\mathcal{T}}$ as the directed multigraph with node set \mathcal{I} and an edge between two nodes for each interaction between the corresponding individuals at a timestamp in the time period \mathcal{T} . Each edge is endowed with additional data m describing the interaction. For example, if S is a mobile operator, m would be the timestamp, the type of interaction (i.e., call or text), its direction (i.e., which party initiated it), and the duration for calls (see Fig. 1). If S is a close-proximity app, m would be the timestamp and the strength of the signal. We denote by k -hop neighbor of a node $v \in \mathcal{I}$ any node $w \in \mathcal{I}$ such that the shortest path between v and w in $G_{\mathcal{T}}$ is of length k . Given a time period \mathcal{T} , $i \in \mathcal{I}$ an individual and $k = 1, 2, \dots$, we define the k -hop Individual Interaction Graph (k -IIG) $G_{i,\mathcal{T}}^k$ as the subgraph induced in $G_{\mathcal{T}}$ by the set of nodes situated on paths of length at most k starting at node i , excluding interactions between the k -hop neighbors themselves. We denote

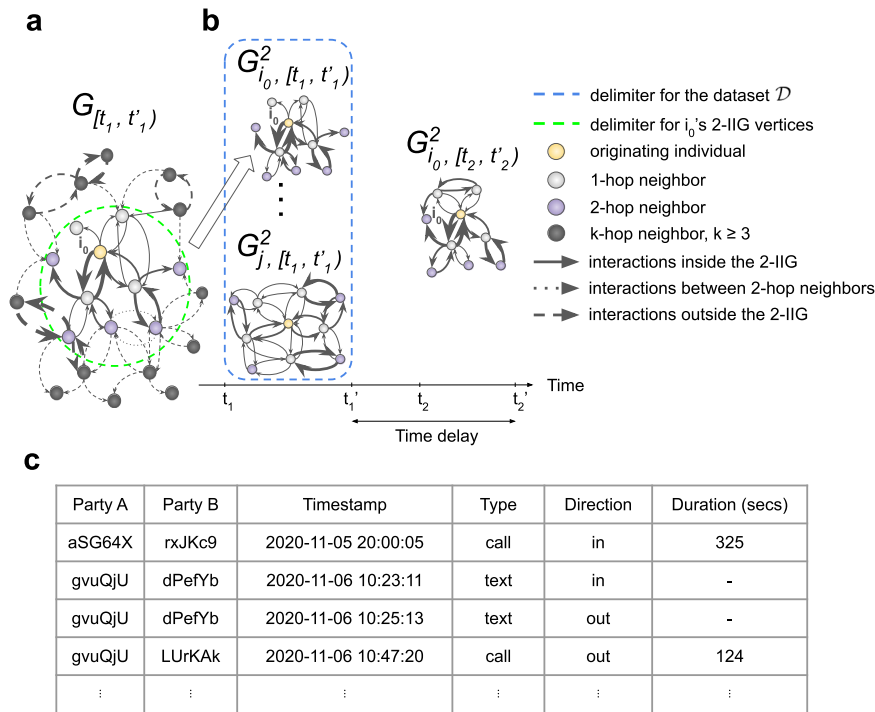


Fig. 1 Setup of the behavioral profiling attack. **a** An example of a 2-IIG is highlighted in the larger graph it comes from. The vertices of the 2-IIG (inside the dashed green circle) are respectively the originating individual (in yellow), 1-hop neighbors (in gray), and the 2-hop neighbors (in purple). In solid lines are the edges that are part of the 2-IIG: all the edges between the originating and 1-hop neighbors; between the 1-hop neighbors; and between 1-hop and 2-hop neighbors, but excluding those between 2-hop neighbors (dotted lines). Dashed lines are all the other edges. For simplicity, all edges are shown as a single directed edge of thickness proportional to the total number of interactions. **b** The data available to the attacker consist of (left) 2-IIGs coming from the time period $[t_1, t'_1]$, usually as part of an anonymized dataset, and (right) auxiliary 2-IIG data about a target individual A ($G^2_{i_0, [t_2, t'_2]}$). While we here display auxiliary data coming from a later period in time, our attack applies equally to cases where the auxiliary data comes from an earlier time period. **c** An example of mobile phone interaction data. Each interaction contains the pseudonyms of the parties A and B, timestamp, type of interactions, direction (equal to “out” if A initiated it, “in” otherwise), and the duration for calls. In this example, the person identified by “gvuQjU” received a text from another person, identified by “dPefYb”, to whom the former responded 2 min later. After 22 min, “gvuQjU” called another individual, identified by “LURkAk”, for a duration of 124 s.

by i the originating individual of k -IIG $G^k_{i, \mathcal{T}}$. Figure 1 shows an example of a 2-IIG.

Our attack model assumes (see Fig. 1) that a malicious agent, the attacker, has access to (1) a dataset $\mathcal{D} = \{G^k_{i, [t_1, t'_1]} : i \in \mathcal{I}'\}$ consisting of the k -IIGs of people in $\mathcal{I}' \subset \mathcal{I}$ from time period $\mathcal{T}_1 = [t_1, t'_1]$, as well as to (2) auxiliary data $G^k_{i_0, [t_2, t'_2]}$ consisting in the k -IIG of a known target individual $i_0 \in \mathcal{I}'$, coming from a disjoint time period $\mathcal{T}_2 = [t_2, t'_2]$ (i.e., $t'_1 \leq t_2$ or $t'_2 \leq t_1$). We further assume that the attacker knows, for each k -IIG, which node is at the center of the k -IIG (originating node), and that the k -IIGs are pseudonymized, meaning that a node will have a different pseudonym in each graph it appears in. The attacker’s goal is to find the target i_0 in \mathcal{D} , i.e., find the $G^k_{i, [t_1, t'_1]} \in \mathcal{D}$ such that $i = i_0$. If successful, the attacker is said to have identified i_0 and is able to retrieve all their interactions from time period $[t_1, t'_1]$. We denote by time delay the quantity $D = t'_2 - t'_1$. We refer the reader to the section “Discussion” for examples.

Model. Our k -IIG-based Behavioral Profiling approach (BP-IIG) first computes a time-dependent profile of an individual in the form of a vector representation (embedding). We apply a neural network to people’s k -IIGs before identifying them using the nearest neighbor in the embedding space.

One of the key challenges for using deep learning in such a setting is that, unlike images or acoustic signals, graphs have a non-Euclidean structure. Recently, generalizations of deep

learning architectures (in particular, convolutional neural networks) have been proposed for graph-structured data^{61,63–65}, with successful applications to biology^{66–71}, medicine⁷², and social network analysis^{6,66}.

To compute the time-dependent profile embedding of individual i , we aggregate the interaction data from their k -IIG $G^k_{i, \mathcal{T}}$, using the nodes’ bandicoot features⁷³ (see Supplementary Tables 1 and 2 and the Supplementary Methods) and by employing a multi-layer graph neural network ($k \geq 2$, see the “Methods” section) of the form:

$$\mathbf{h}_i^{(s)} = \xi^{(s)} \left(\left[\mathbf{h}_i^{(s-1)}, \sum_{j \in \mathcal{N}(i)} \alpha_j^{(s)} \mathbf{h}_j^{(s-1)} \right] \right) \quad (1)$$

$$\alpha_j^{(s)} = \frac{\alpha^{(s)}(\mathbf{h}_i^{(s-1)}, \mathbf{h}_j^{(s-1)})}{\sum_{l \in \mathcal{N}(i)} \alpha^{(s)}(\mathbf{h}_i^{(s-1)}, \mathbf{h}_l^{(s-1)})} \quad (2)$$

where the output $\mathbf{h}_i^{(s-1)}$ of layer $s-1$ is passed as the input to layer $s = 1, \dots, S$. For each layer $1 \leq s \leq S$, $\xi^{(s)}$ is a non-linear parametric function implemented as a multi-layer perceptron (MLP) with one hidden layer, followed by \mathbb{L}_2 -normalization. Finally, $\alpha^{(s)}$ denotes the attention weight computed as a nonlinear parametrized function of the features of node i and its neighbor $j \in \mathcal{N}(i)$. The neural attention mechanism, previously shown to improve performance in tasks such as object recognition⁷⁴ and machine translation⁷⁵, has been adapted for graph inputs by aggregating a

node’s neighborhood features via a weighted average over the features of the neighbors⁶². The attention weights are potentially different for distinct neighbors and are optimized for a specific learning task.

The network is applied to the input node-wise features $\mathbf{h}_i^{(0)}$ and its output $\mathbf{h}_i^{(s)} = \mathbf{h}(G_{i,T}^k; \Theta)$ is used as the embedding of individual i , with Θ denoting the network parameters of $\xi^{(s)}$ and $\alpha^{(s)}$ optimized during training.

The neural network is trained to optimize the matching accuracy, using the triplet loss⁷⁶, which optimizes the profile embeddings of the same individual at different time periods (positive pair) to be closer to each other than to those of different individuals at any time period (negative pair). A triplet of k -IIGs ($G_{i,T}^k, G_{i',T'}^k, G_{i',T''}^k$) contains data from two individuals $i \neq i'$, such that there are two k -IIGs from i , coming from time periods that are not equal, but could be overlapping $T \neq T'$, and a k -IIG from i' from a time period T'' (not necessarily different from T or T'). Let $\mathbf{h}(\Theta) = \mathbf{h}(G_{i,T}^k; \Theta)$, $\mathbf{h}^+(\Theta) = \mathbf{h}(G_{i,T'}^k; \Theta)$ and $\mathbf{h}^-(\Theta) = \mathbf{h}(G_{i',T''}^k; \Theta)$ denote the respective embeddings. The triplet loss

$$\ell(\Theta) = \max(0, \|\mathbf{h}(\Theta) - \mathbf{h}^+(\Theta)\|_2 - \|\mathbf{h}(\Theta) - \mathbf{h}^-(\Theta)\|_2 + \lambda) \quad (3)$$

tries to ensure that the profiles (\mathbf{h}, \mathbf{h}^+) of the positive pair (i.e., the pair of profiles constructed from interaction data of the same individual, but different time periods) are closer than those (\mathbf{h}, \mathbf{h}^-) of the negative pair (i.e., the pair of profiles constructed from i and another individual’s interaction data in possibly, but not necessarily, different time periods T and T'') by at least a margin λ . We average the triplet loss over a training set of positive and negative pairs and minimize it w.r.t. the network parameters Θ . The optimal parameters Θ^* obtained as the result of training are then used for the attack.

The attacker trains the embedding network on data from the dataset \mathcal{D} (see the “Methods” section). To identify the target individual i_0 in \mathcal{I}' , the attacker computes the Euclidean distance $d_{i_0,j} = \|\mathbf{h}(G_{i_0,T_2}^k; \Theta^*) - \mathbf{h}(G_{j,T_1}^k; \Theta^*)\|_2$ between the profile of i_0 from target time period $\mathcal{T}'_2 \subset \mathcal{T}_2$ and the profiles of all the individuals $j \in \mathcal{D}$ from a reference time period $\mathcal{T}'_1 \subset \mathcal{T}_1$ of same length as \mathcal{T}'_2 . If the candidate with the smallest distance is (resp. R candidates with the smallest distance contains) the target individual (i.e., $i_0 \in \{j_1, \dots, j_R\}$), we say that we have correctly identified i (resp. within rank R).

Mobile phone interaction data. We use a mobile phone interaction dataset composed of the 3-IIGs of $N = 43,606$ subscribers of a mobile carrier collected over a period of $T = 35$ consecutive weeks $\mathcal{T} = \mathcal{W}_1 \cup \dots \cup \mathcal{W}_T := \mathcal{W}_{1:T}$, where $\mathcal{W}_n = [t_n, t_{n+1})$ denotes the n th week, with $1 \leq n \leq T$ and t_{n+1} and t_n differing by one week. The interaction data contain the pseudonyms of the interacting parties, timestamp, as well as the type of interaction (call or text), the direction of the interaction, and the duration of calls. We here consider the auxiliary profiling information available to the attacker to be the k -IIG of the target individual from a week $\mathcal{T}_2 \in \{\mathcal{W}_{T'+1}, \dots, \mathcal{W}_T\}$ and the anonymous dataset to be the k -IIGs of all the N people from the first $T' = 15$ weeks of data ($\mathcal{T}_1 = \mathcal{W}_{1:T'}$). We report the probability of identification within rank R , defined as the fraction of people among the N subscribers who are correctly identified within rank R (averaged over 10 runs).

Figure 2 shows that our model correctly identifies people $p_{k=2} = 52.4\%$ of the time in a dataset of 43.6k people with $k = 2$ i.e. when the attacker has access to an individual’s interactions as well as the interactions of their contacts here with a time delay of a week. It also shows the probability p of identification of a target individual within the top R matches. Our model is able to rank

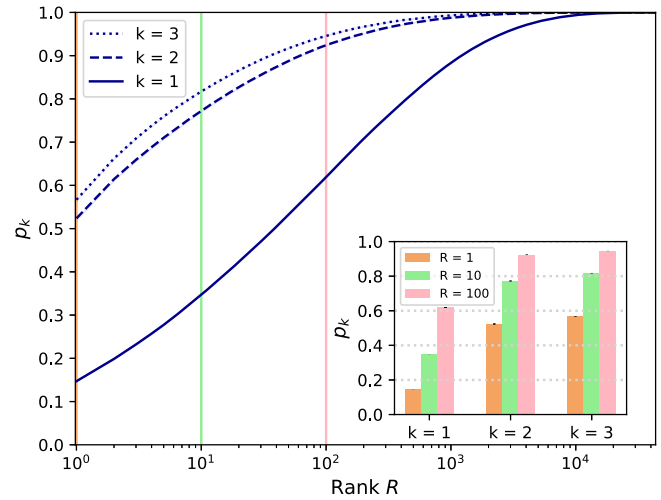


Fig. 2 Probability of identification. For each $k \in \{1, 2, 3\}$, we plot p_k , the probability of identification within rank $R \in \{1, \dots, 43, 606\}$ when the time delay is $D = 1$ week, with the 95% confidence interval shown in light blue. (Inset) shows the probability of identification for ranks 1, 10, and 100, with error bars for the 95% confidence interval. Our model correctly identifies people 52.4% of the time for $k = 2$. The probability of correct identification is still high at $p_{k=1} = 14.7\%$ for $k = 1$ and slightly increases $p_{k=3} = 56.7\%$ when k increases from 2 to 3. Our model ranks the correct candidate among the top 10 predictions $p_{k=2} = 77.2\%$ of the time and among the top 100 predictions $p_{k=2} = 92.4\%$ of the time for $k = 2$.

the correct person among the top 10 candidates $p_{k=2} = 77.2\%$ of the time and among the top 100 candidates, $p_{k=2} = 92.4\%$ of the time.

When $k = 1$, i.e., when the attacker has only access to the individual’s direct interactions, our model is still able to identify people $p_{k=1} = 14.7\%$ of the time. While having access to the 2-hop information helps, our model still performs much better than random for $k = 1$. The probability of identifying the correct person among the top 10 candidates (rank 10) is $p_{k=1} = 34.7\%$ while the rank 100 probability is $p_{k=1} = 61.9\%$, respectively. Interestingly, having access to information beyond the target’s direct contacts ($k = 3$) only marginally increases the probability of correct identification $p_{k=3} = 56.7\%$ (a 7.9% increase w.r.t. $k = 2$). Higher ranks probabilities similarly increase to $p_{k=3} = 81.7\%$ and $p_{k=3} = 94.6\%$, respectively, a 5.8% and a 2.4% increase. On the one hand, this marginal increase could be due to the fairly large number of nodes reached with $k = 3$ (121.5 ± 48.8 for $k = 3$ vs. 17.3 ± 13.4 for $k = 2$) thereby limiting the usefulness of data from larger k (see Supplementary Note 1). On the other hand, this could also be due to our particular choice of architecture. In particular, while we downsampled the simplified k -IIG to contain no more than $\tau = 200$ nodes for $k = 3$ (see the Supplementary Methods), the graph neural network architecture might still suffer from over smoothing. Given that new architectures could be developed to leverage information coming from the 3-IIG specifically, from a privacy perspective, our results are thus only a lower bound on the risk of re-identification.

The accuracy of our model is likely to decrease as time passes: people change behavior, make new friends, and lose contact with others. Figure 3 shows that, despite this, the probability of correct identification only slowly decreases with the time delay $D = t'_2 - t'_1$ (see the section “Setup”). Even after 20 weeks, our model still correctly identifies people $p_{k=2} = 24.3\%$ of the time when $k = 2$. This suggests that the profiles our model extracts from the data capture key behavioral features of individuals. The

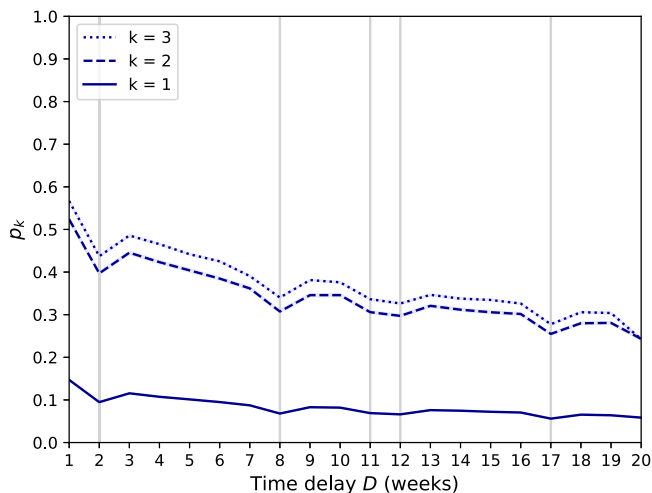


Fig. 3 Probability of identification when the time delay increases. We plot p_k , the probability of identification within rank 1 for $k \in \{1, 2, 3\}$ when the time delay between the dataset and the attacker’s auxiliary information is equal to D weeks. The auxiliary information is one week long. The 95% confidence interval is shown in light blue. The vertical gray lines correspond to holidays. While p_k decreases slowly with the time delay, our model correctly identifies people $p_{k=2} = 24.3\%$ of the time even after 20 weeks ($k=2$). Even for $k=1$, the probability after 20 weeks is as high as $p_{k=1} = 5.8\%$. The probability of identification decreases as $y = p_k(D = 1) - \alpha_k \times (D - 1)$, with $\alpha_1 = -0.006$, $\alpha_2 = -0.017$ and $\alpha_3 = -0.018$.

probability of identification decreases similarly slowly with time for $k=3$ and $k=1$.

Interestingly, Fig. 3 shows that the probability of identification (p_k) visibly decreases when the time delay is 8, 11, 12, and 17 weeks, respectively. In a post-hoc analysis, we found that they all correspond to weeks containing a national holiday. This further suggests that our model captures a person’s routine weekly behavior, both weekdays and weekends, and consequently loses some accuracy when a user’s behavior changes in response to external events.

We have so far assumed that the attacker has access to a week of a target individual’s data, i.e., their auxiliary information is the target individual’s k -IIG from one week. In practice, an attacker might often have access to more weeks of data from an individual. In the D4D challenge, data were for instance re-pseudonymized every 2 weeks⁷⁷ while a company wanting to archive transactional data might decide to pseudonymize and archive it on a monthly basis. To simply evaluate the extent to which more auxiliary data increase accuracy, we combine the predictions from growing sequences of target weeks used as auxiliary data. For $1 \leq L \leq T - T'$ (L denotes the number of weeks in the auxiliary data or \mathcal{T}_2), we combine the predictions from the $T' + 1, \dots, (T' + L)$ th target weeks using a majority vote: the candidate that was ranked first most of the time is the final prediction. The tie-breaks are decided by the lowest total distance between the target individual and the highest-ranked candidate (see Supplementary Note 2).

Figure 4 shows how having auxiliary data over several weeks further improves the performance of the attack. For $k=2$, the probability of correct identification increases from $p_{k=2} = 52.4\%$ with one week of auxiliary data to $p_{k=2} = 66.0\%$ with $L = 16$ weeks. Interestingly, the probability of correct identification for all values of k increases fast and then plateaus around $L = 8$, even slightly decreasing after $L = 16$ and $L = 15$ for $k=2$ and $k=3$, respectively. Despite having access to more data, the attack is less accurate for increasing time delay. While this might

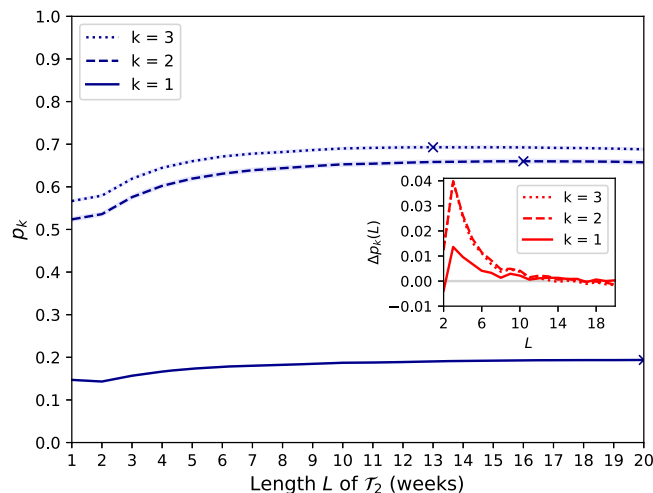


Fig. 4 Probability of identification for increasing time period length of auxiliary data. For each $k \in \{1, 2, 3\}$, we plot p_k , the probability of correct identification ($R = 1$) when the attacker’s auxiliary data \mathcal{T}_2 consist of L weeks, $1 \leq L \leq 20$ (the largest value for each k is marked). The 95% confidence interval is shown in light blue. (Inset) shows the difference quotient $\Delta p_k(L) = p_k(L) - p_k(L - 1)$ for $2 \leq L \leq 20$. The probability of correct identification increases fast before plateauing around $L = 8$ weeks for all values of k , even slightly decreasing after $L = 16$ and $L = 15$ for $k=2$ and $k=3$, respectively. The largest values are $p_{k=1} = 19.4\%$ at $L = 20$, $p_{k=2} = 66.0\%$ at $L = 16$, and $p_{k=3} = 69.3\%$ at $L = 13$. This shows that having more auxiliary data further improves the performance of the attack, although data that are more distant in time seem less useful than closer ones, even slightly detrimental.

seem surprising at first, we hypothesize this to be due to small changes in people’s behavior over time. This makes auxiliary data that are more distant in time less useful than closer ones and sometimes slightly detrimental. The maximum probability for $k=2$ is at $L = 16$ weeks ($p_{k=2} = 66.0\%$) and for $k=1$ and $k=3$ at $L = 20$ ($p_{k=1} = 19.4\%$) and $L = 13$ ($p_{k=3} = 69.3\%$), respectively. Finally, we show that the accuracy of our attack only decreases slowly with the size of the dataset size (see Supplementary Note 3 and Supplementary Fig. 3).

We finally perform a post-hoc analysis to better understand who are the people that our model identifies correctly. Figure 5 shows (in blue) in how many weeks a person is correctly identified by our attack, each time using a single week of auxiliary data target weeks (weeks $T' + 1, \dots, T$ of the mobile phone dataset). For instance, for $k=2$, 86.8% of people are correctly identified by our model at least once (5% of the 20 target weeks). We compare this with a naïve model in which individuals are identified independently in each week with the same probability as our attack, and independently from one another. In the latter setting, the number of weeks when a person is correctly identified follows a Poisson binomial distribution defined as the probability distribution of $B := \sum_{l=T'+1}^T B_l$ with $B_l \sim \text{Bernoulli}(p_l)$, where p_l denotes the probability of identification in target week l using our attack (see the Supplementary Note 4). We can see that our attack identifies some people in many more weeks than expected. For $k=2$, the people we identify more often than expected are correctly identified in at least 40% of the weeks. The two curves cross one another at 20% and 45% for $k=1$ and $k=3$ respectively. In all the other initializations of our attack and every $k \in \{1, 2, 3\}$, the lowest abscissa value where our approach outperforms the baseline is the same.

Figure 6 suggests that, when holding all other features constant, individuals with more interactions, or a well-balanced interaction

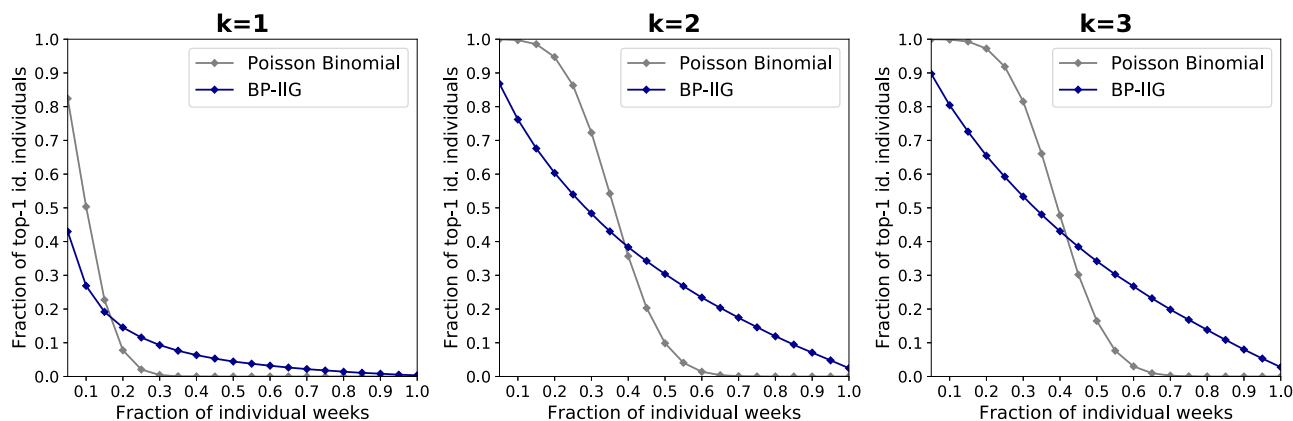


Fig. 5 Fraction of identified people vs. fraction of individual weeks. For each $k \in \{1, 2, 3\}$, we plot the fraction of people that are identified in at least a given fraction of individual weeks, using our model (in blue) and according to a Poisson binomial distribution (in gray, averaged over 100 trials). Our attack identifies 38.4% (resp. 14.5% and 38.5%) of the people more often than expected for $k = 2$ (resp. for $k = 1$ and $k = 3$).

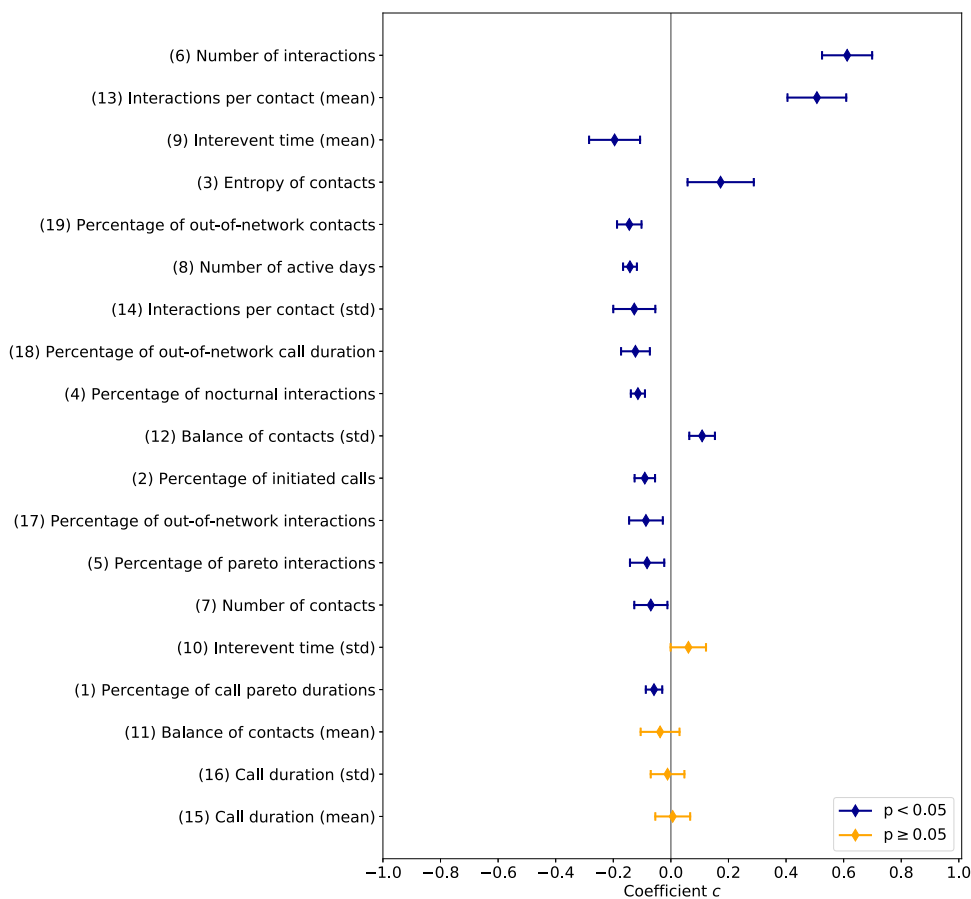


Fig. 6 Coefficients of logistic regression for individual identifiability. For each feature, we plot the coefficient c (with the 95% confidence interval) of a logistic regression classifier with whether a person is more or less identifiable than expected as the dependent variable. Features are ordered decreasingly from top to bottom according to the absolute value of c . When holding all other features constant, these results suggest that having more interactions and a well-balanced interaction graph makes individuals more identifiable.

graph are more identifiable. Using logistic regression, we study in a post-hoc analysis what distinguishes the people our model identifies more often and less often than expected for $k = 2$. Supplementary Table 3 shows the bandicoot features used in this analysis. The largest coefficients (in absolute value), both as individual predictors (see Supplementary Fig. 4) and taken together, are the number of interactions, the mean number of interactions per contact ($c > 0$), and the mean interevent time,

(i.e., time elapsed between consecutive interactions) ($c < 0$). Interestingly, a person’s call duration (both mean and standard deviation) seems to have no impact ($p \geq 0.05$) on identifiability. While the standard deviations of all summary distributions are highly correlated with their mean ($\rho > 0.7$, see Supplementary Fig. 5), they can still be informative even when other features are accounted for, e.g., the standard deviation of the number of interactions per contact. Last, we note that all other features being

the same, the lower a person's number of active days, the more likely they are to be identified, with similar findings for the percentage of nocturnal or out-of-network activity. A more detailed analysis of the logistic regression results and pairwise feature correlations is provided in Supplementary Note 4. While our findings suggest the possible influence of the various behavioral features on identification, a causal analysis is beyond the scope of this paper.

Bluetooth close-proximity data. To prevent the spread of COVID-19, governments and companies around the world have been developing and releasing a number of contact tracing apps. Contact tracing apps use Bluetooth to collect close-proximity data between users. If a user becomes infected, they upload to a server data allowing their contacts to be informed that they might have been infected. In the centralized model, application users typically upload the temporary pseudonyms of their contacts^{78,79}. In the decentralized model, they upload data about themselves, typically cryptographic keys, which their contacts can use to deduce that they might have been infected^{80–83}. In another (“hybrid”) system, users upload their encounter keys (corresponding to a pair of user identifiers) instead⁸⁴. Numerous application designs based on these protocols have been proposed and are under active development.

Our attack is, to the best of our knowledge, the first to show how mitigation strategies relying on changing pseudonyms of both the person and of all of their contacts could fail to adequately protect people's privacy. While it does not target a specific application, protocol, or type of protocol (centralized, decentralized, or hybrid), it could form an effective basis for an attack against any system where an attacker has access to a user's social graph over two or more time periods. This could be by design in a centralized system (e.g., the UK's NHSX app reportedly plans to change keys every 24 h⁷⁸) or the results of extra data collection in a decentralized system (e.g., the Belgian system reportedly collects the number of encounters with infected users and, for each encounter, the number of days elapsed since the reported contamination of the other user⁸⁵). While the specifications for the reporting of data for epidemiological purposes are currently under discussion, they are likely to include part or all of the infected user's social graph.

We evaluate the effectiveness of our attack using a real-world Bluetooth close-proximity network of 587 university students over 4 weeks¹¹. Our interaction data consist of the identifiers of the parties, the interaction timestamp and the received signal strength indication (RSSI), a proxy for the distance between devices. This is the data typically captured by contact tracing apps⁷⁸.

Figure 7 shows that for $k = 1$ our approach is able to identify target individuals $p_{k=1} = 26.4\%$ of the time among the 587 people. Out of 10 people ($R = 10$), it is able to identify the right person $p_{k=1} = 60.1\%$ of the time. While our dataset is too small to evaluate for larger values of k , we expect the results to further increase when more information is available.

Taken together, our results provide strong evidence of the urgent need to consider profiling attacks when evaluating whether systems, protocols, or datasets satisfy Article 29 WP's definition of anonymization³¹. In particular, they show how people's interaction patterns online and offline remain identifiable across long periods of time allowing an attacker to link together data coming from disjoint time periods with high accuracy even in large datasets. Our results challenge current data retention practices and, in the context of the recent COVID-19 pandemic, whether some of the collected data would satisfy the Article 29 linkability criteria. They finally further question the policy

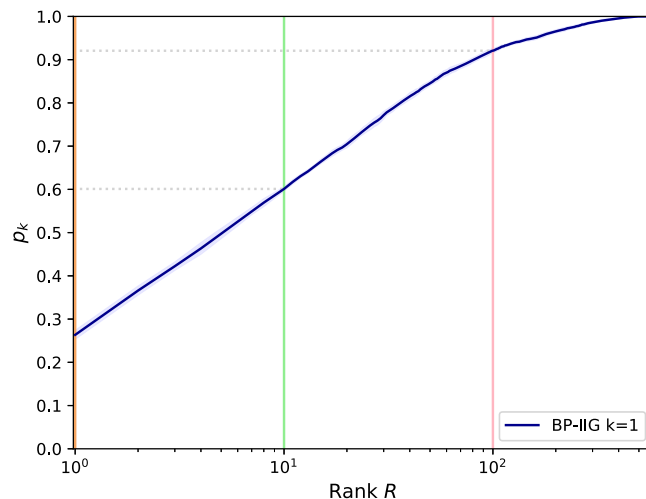


Fig. 7 Probability of identification in a bluetooth close-proximity network.

We plot $p_{k=1}$, the probability of identification within rank R for $k = 1$. The 95% confidence interval is shown in light blue. Our method correctly identifies people $p_{k=1} = 26.4\%$ of the time based on their 1-IIGs. Out of 10 people ($R = 10$), it is able to identify the right person $p_{k=1} = 60.1\%$ of the time.

relevance of de-identification techniques⁸⁶ and emphasize the need to rethink our approaches to safely use non-personal data. In particular, legal and access control mechanisms are necessary to protect data retained in pseudonymized format, and privacy engineering solutions such as query- and question-and-answer-based systems, local DP mechanisms, or secure-multiparty computation could be deployed to help use data anonymously⁸⁷.

Discussion

In this paper, we propose a new behavioral profiling attack model exploiting the stability over time of people's k -hop interaction networks. We evaluate its effectiveness on two real-world offline and online interaction datasets and show the risk of identification to be high.

We first compare our attack to previous work from 2014⁴⁹, the only attack in the literature developed for user linkage across call graphs in the context of the D4D challenges (hereafter: ShDa). The method uses a random forest classifier trained on hand-engineered node pair features representative of the nodes' 2 or 3-hop neighborhoods. The node pair features are pairwise combinations of individual node features consisting of the histogram of each node's 1-hop or 2-hop neighbors' degrees. We reimplement their attack for matching nodes from two networks based on nodes' k -hop neighborhood features, $k \leq 3$, in each network, respectively, and compare their results to ours. For a fair comparison, we convert our attack, which computes a target individual's match by distance comparison with a list of candidates, into their setup: a binary classifier predicting as positive any pair with distance lower than a threshold (see Supplementary Note 5).

Figure 8 shows that our approach (BP-IIG, blue line) vastly outperforms previous work (ShDa, solid green line) making profiling attacks a real risk. We report the receiver operator characteristic (ROC) curve and area under the curve (AUC) on the binary classification task for $k = 2$, showing show our approach achieves, on their task and for a false positive rate of 0.05, a true positive rate of 0.99 (AUC = 0.998) vs. 0.36 for ShDa (AUC = 0.868). Our method still outperforms ShDa when we add to it our behavioral features (ShDA + BF, green dashed line) which result in a true positive rate of 0.82 for a false positive rate of 0.05. We refer the reader to Supplementary Fig. 6 for results for

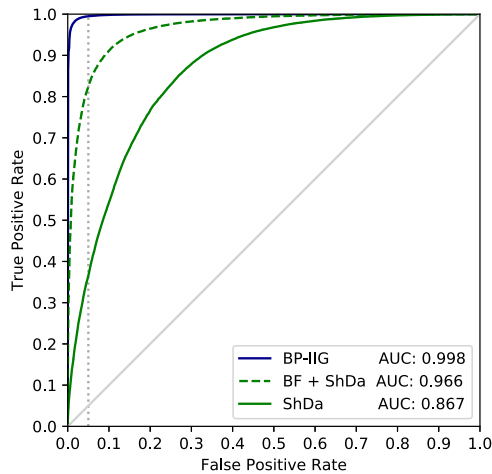


Fig. 8 Receiving operator characteristic (ROC) curves with area under the curve (AUC) score for the node pair classification task for $k = 2$. The performance of a random classifier is shown in solid gray. Even for their task, our method (blue line) vastly outperforms both ShDa (solid green line) and its improved version ShDa + BF (dashed green line). For a false positive rate of 0.05, our method achieves a true positive rate of 0.99 vs. 0.36 for ShDa and 0.82 for ShDa + BF (vertical dotted gray line). As shown in Supplementary Fig. 7, both ShDa and ShDa + BF perform poorly when it comes to correctly identifying a person.

other values of k . More importantly, Supplementary Fig. 7 shows how our approaches strongly outperform ShDa on the task of interest: the probability p_k of correctly identifying a person. Here, ShDa alone only achieves a $p_{k=2} = 0.3\%$ versus $p_{k=2} = 52.4\%$ for our attack. Even the improved version, ShDa + BF, only achieves a low $p_{k=2} = 8.3\%$. Our approach further improves on other baselines (see Supplementary Note 5).

We further validate that our attack generalizes by examining its performances when testing is performed on a set disjoint from the training set in the identities of the individuals, time periods used, or both, as illustrated in Supplementary Fig. 8. Our attack performs similarly across the three scenarios for all values of $k \in \{1, 2, 3\}$, as shown in Supplementary Table 5. Our attack is equally able to identify people unseen during training in time periods also unseen during training ($p_{k=2} = 61.5\%$) as in cases when the same people ($p_{k=2} = 62.2\%$) or time periods ($p_{k=2} = 60.5\%$) used in testing are seen during training. We observe similar results for $k = 1$ and $k = 3$ (see Supplementary Note 6).

While the attack model is general (see Setup), we have throughout the paper assumed that the auxiliary information comes from a time period posterior to the dataset \mathcal{D} ($t'_1 < t'_2$). Using our BP-IIG ($k = 2$) approach, we compared the performance of a model trained on 9 consecutive weeks of data and tested on the following 9 weeks, with that of a model trained on the last 9 weeks and tested on the first 9 weeks. The two models gave the same performance ($p = 0.58$, see Supplementary Note 7). This confirms the generality of our model.

We here focus on a general attack model which we use to show how both mobile phone and bluetooth interaction data are identifiable across long periods of time. While we do not wish to emphasize specific attack scenarios, examples could include data collectors pseudonymizing interaction data monthly as part of their data retention policy; poorly designed centralized contact tracing apps relying on frequent re-pseudonymization to protect user's privacy; or the behavioral identification of a phone through e.g. their messaging pattern. The attacker could also be a law enforcement agency with, e.g., the Patriot Act giving intelligence

agencies access to the 3-hop graphs of suspects (later restricted to 2-hop under the 2015 USA Freedom Act)⁸⁸.

Our attack model uses a definition of the k -IIG that excludes interactions between the k -hop neighbors, as already done in the past for mobile call graphs⁷⁷. We consider this to be a realistic assumption e.g. for $k = 1$ when the attacker's auxiliary information could come from the target's mobile. In the context of contact tracing, the attacker would have access to the log of the target's interactions with their contacts but would not have any information on the interactions between its contacts. This assumption makes our results a lower bound of what could be achieved with more information.

While we assume, again in line with previous practices^{49,77}, that pseudonyms are identical over time for nodes in k -IIGs of the same individual, this is not a requirement of our approach. Repseudonymization of nodes over time might be used, for example, to avoid direct access to an individual's interactions over a long period of time. Our approach would still work even if the dataset \mathcal{D} consisted of weekly k -IIGs with different pseudonyms for the same node appearing in two weekly k -IIGs of the same person, so long as the attacker knows the identity of the originating individual in each weekly k -IIG. For $k = 1$, this is due to the approach relying on the originating individual's behavioral features. For $k \geq 2$, the graph attention network used is invariant to nodes' ordering, but the originating individual's identity is needed for computing the k -IIG's final embedding.

Methods

Overview of the attack. We assume that the dataset and the auxiliary data come from disjoint time periods \mathcal{T}_1 and \mathcal{T}_2 , respectively. The attack is based on comparing an individual's weekly profile extracted from time period \mathcal{T}_2 to the weekly profiles of everyone in the dataset, constructed from their respective weekly k -IIGs in \mathcal{T}_1 . The attack thus exploits the weekly patterns in human behavior (e.g., weekdays and weekend). We assume \mathcal{T}_1 and \mathcal{T}_2 to be at least one week long. The attacker splits the k -IIGs from $\mathcal{D} = \{G_{i,\mathcal{T}_1}^k : i \in \mathcal{I}'\}$ by weeks to obtain $\{G_{i,\mathcal{W}_t}^k : i \in \mathcal{I}', 1 \leq t \leq T'\}$, where $\mathcal{T}_1 = \mathcal{W}_1 \cup \dots \cup \mathcal{W}_{T'}$. From \mathcal{T}_2 , the attacker extracts one target week $\mathcal{T}'_2 \subset \mathcal{T}_2$ about the target individual i_0 . The attacker then singles out the most recent week in \mathcal{T}_1 , reference week $\mathcal{T}'_1 \subset \mathcal{T}_1$, to be used for the identification. The remaining data in \mathcal{T}_1 are used to train the profiles of k -IIGs.

Preprocessing of a k -IIG. The attacker extracts behavioral features at the weekly level, then simplifies each weekly k -IIG to a simple graph that can be mapped to an embedding using graph neural networks and optimized for identification.

We use bandicoot⁷³, an open-source Python library to compute a set of behavioral features from an individual's list of interactions. Bandicoot has been used to predict people's personality²⁰, making it a suitable choice for identification. bandicoot takes as input an individual's list of interactions, consisting of the other party's unique identifier, the interaction timestamp, type (call or text), direction (in or out), and duration (if a call). The features range from simple aggregated features, e.g., the number of voice and text contacts, to more sophisticated statistics, e.g., the percentage of an individual's contacts that account for 80% of their interactions. For the Bluetooth close-proximity data, we set the type to call, the direction to out, and the call duration to the negative RSSI. Supplementary Tables 1 and 2 list the features used in this paper for the mobile phone dataset and the Bluetooth close-proximity dataset, respectively.

Using bandicoot, the attacker extracts a set of behavioral features for all nodes in a weekly k -IIG with outdegree ≥ 1 that are at most $k-1$ hops away from the originating individual. In practice, the positive outdegree is a proxy for a node being a subscriber to service S . To these features the attacker adds estimates of the percentage of out of network call, texts, call durations, and contacts based on the information available in k -IIG. The attacker further removes the featureless nodes from the k -IIG and collapses all directed edges between two remaining nodes into a single directed edge of the same direction. The attacker thus simplifies the k -IIG $G_{i,\mathcal{T}}^k = (V, E)$ to obtain the simplified k -IIG $\tilde{G}_{i,\mathcal{T}}^k = (\tilde{V}, \tilde{E})$, a simple graph with $\tilde{V} = \{v \in V : v \text{ is on a path of length at most } k-1 \text{ from node } i\} \cap \{v \in V : \exists w \in V \text{ with } (v, w, m) \in E\}$ and $\tilde{E} = \{e = (v, w) \in \tilde{V} \times \tilde{V} : v \neq w \wedge \exists (v, w, m) \in E\}$ (see the Supplementary Methods).

Embedding of k -IIG. Our k -IIG-based Behavioral Profiling approach (BP-IIG) first computes a time-dependent profile of an individual in the form of a vector representation (embedding) by aggregating the features in $\tilde{G}_{i,\mathcal{T}}^k$ using graph neural

networks with attention, similarly to the GraphSAGE architecture⁶⁶, but using attention weights⁶², as described in Supplementary Alg. 1. Supplementary Fig. 2B illustrates the model and Supplementary Note 8 shows an analysis of the attention weights. Differently from GraphSAGE, the architecture uses an MLP with a hidden layer instead of a single fully connected layer after each concatenation between the features of the node originating the simplified k -IIG and the weighted average of its neighbors' features. The output of the MLP layer is \mathbb{L}_2 -normalized.

Triplet sampling procedure. The embeddings are optimized for identification using the triplet loss⁷⁶ with a triplet sampling procedure designed with the goal of separating the profiles of different individuals in the embedding space. A triplet is composed of an anchor, a positive, and a negative example. The anchor and positive examples are two instances of the same individual, while the negative example is an instance of a different individual. For a given batch size B , the triplet sampling procedure works as follows: (1) one week i is sampled uniformly at random among the training weeks, (2) B individuals are sampled from \mathcal{T} and their k -IIGs in week i are used as anchor examples, while their k -IIGs in weeks $i-1$ and $i+1$ (modulo the number of training weeks) are used as positive examples and (3) for each anchor, a negative example is selected via mini-batch moderate negative sampling⁸⁹. For step 3, all k -IIGs in weeks $i-1$, i and $i+1$ coming from the other $B-1$ individuals in the batch are considered as candidate negative examples. In practice, each mini-batch contains $2B$ triplets, because at step 2) two different positive examples are considered for each anchor. Mini-batch gradient descent is used for the optimization. An epoch is defined as a full pass over at least one anchor example of each individual in \mathcal{T} . As described above, the attacker splits the dataset to obtain $T' \times |\mathcal{T}'|$ k -IIGs as follows: $\{G_{i,W_t}^k : i \in \mathcal{T}', 1 \leq t \leq T'\}$, with $T'k$ -IIGs per individual in \mathcal{T}' . Data from $P \leq T'$ weeks are used to train the model. There are, therefore, by construction, exactly P weekly k -IIG instances available for the triplet sampling procedure for each individual in \mathcal{T}' .

Training setup. In the mobile phone dataset, data from enough weeks are available, so the attacker uses disjoint weeks for training: $\mathcal{T}_1 := \mathcal{W}_1 \cup \dots \cup \mathcal{W}_{T'}$, with $\mathcal{W}_1, \dots, \mathcal{W}_{T'}$ disjoint and ordered increasingly. Week $\mathcal{W}_{T'}$ is used as reference week \mathcal{T}'_1 in the attack. For each $k \in \{1, 2, 3\}$, the attacker selects the best hyperparameters using cross-validation on the weeks $\mathcal{W}_{1:T'-1}$, where each test fold is composed of two consecutive weeks. The first week is used as reference week and the auxiliary data about target individuals come from the second week. With T' being odd, the $(T'-1)/2$ disjoint test folds are defined as $\{(\mathcal{W}_{2i+1}, \mathcal{W}_{2i+2}), 0 \leq i < (T'-1)/2\}$. For each fold, the previous two time periods (modulo $T'-1$) are used as validation weeks for early stopping. The remaining weeks are used for training. Given the best hyperparameter set, the attacker trains the model on data from $\mathcal{W}_{1:T'-3}$, using validation weeks $(\mathcal{W}_{T'-2}, \mathcal{W}_{T'-1})$ for early stopping. For early stopping, the metric used is p_k , the probability of identification within rank 1 on the validation weeks.

In the Bluetooth close-proximity dataset, only 4 weeks, here denoted $\mathcal{T}_1 = \mathcal{W}_1 \cup \dots \cup \mathcal{W}_4 := \mathcal{W}_{1:4}$ are available. For $k=1$, the attacker uses the first two weeks of data for training, the second and third week of data for validation, and results are reported on the third and fourth week of data (i.e., $\mathcal{T}'_1 = \mathcal{W}_3$ and $\mathcal{T}'_2 = \mathcal{T}_2 = \mathcal{W}_4$). In order to increase the number of training samples per individual, the attacker generates 8 overlapping weeks of data from the two training weeks. Because the training data contain a total of 14 days of interactions $d_1 \cup \dots \cup d_{14}$, the attacker generates 8 overlapping weeks $\mathcal{W}'_1, \dots, \mathcal{W}'_8$, with $\mathcal{W}'_i = d_i \cup \dots \cup d_{i+6}$, $1 \leq i \leq 8$.

Data availability

The Bluetooth close-proximity dataset¹¹ is available at <https://doi.org/10.6084/m9.figshare.7267433>. For contractual and privacy reasons, we cannot make the raw mobile phone data available.

Code availability

To limit the risk of nefarious uses we chose—in coordination with ethics reviewers—to not publicly release the code. We will instead make the code available upon request to the corresponding author to researchers in the field for scientific purposes.

Impact statement

We hope our findings will help raise awareness of the risk posed by the identifiability of interaction data. In particular, we hope this will encourage the implementation of security measures and the deployment of privacy-preserving systems when collecting, analyzing, and sharing such data.

Our attack is a general profiling attack against interaction data. While we show our attack to be effective against bluetooth interaction data—the same type of data collected by contact tracing applications—we neither attacked nor considered specific applications or protocols. For the avoidance of doubt, we do not believe our results currently apply to robust privacy-preserving contact tracing protocols such as Google and Apple's Exposure Notification (GAEN).

While the publication of our findings might increase the risk of profiling attacks being used for nefarious purposes, we believe the benefits of these findings being public

knowledge, alongside our decision not to release the code and to delete the models upon publication, means that the benefits largely outweigh the risks in general. First, we believe that deploying an attack similar to ours was already possible as the technologies used (graph attention networks, bandicoot features, etc.) are already well-known in the literature. The publication of our results will instead inform practitioners about the risk and enable them to enact security measures. Second, to limit the reach and possible misuse of our attack in practice, we chose—in coordination with ethics reviewers—not to release the code publicly and to only make it available upon request to researchers in the field for scientific purposes. Third, we considered developing and releasing, alongside our findings, technical defenses. While defenses such as noise addition might mitigate the risk, we were not convinced they would effectively prevent future attacks in general. Worse, they might give a false impression that privacy is preserved. Instead, we believe security measures such as access control and privacy-enhancing systems based on provable guarantees to be the best defenses today against profiling attacks.

Received: 20 June 2020; Accepted: 29 November 2021;

Published online: 25 January 2022

References

- Lazer, D. et al. Computational social science. *Science* **323**, 721–723 (2009).
- Blondel, V., Krings, G. & Thomas, I. Regions and borders of mobile telephony in Belgium and in the Brussels metropolitan zone. *Brussels Stud* **42**, 1–12 (2010).
- Saramäki, J. et al. Persistence of social signatures in human communication. *Proc. Natl Acad. Sci. USA* **111**, 942–947 (2014).
- Bengtsson, L. et al. Using mobile phone data to predict the spatial spread of cholera. *Sci. Rep.* **5**, 8923 (2015).
- Shao, C. et al. The spread of low-credibility content by social bots. *Nat. Commun.* **9**, 1–9 (2018).
- Monti, F., Frasca, F., Eynard, D., Mannion, D. & Bronstein, M. M. Fake news detection on social media using geometric deep learning. ICLR 2019 Workshop on Representation learning on graphs and manifolds (ICLR, 2019).
- Bovet, A. & Makse, H. A. Influence of fake news in Twitter during the 2016 US presidential election. *Nat. Commun.* **10**, 1–14 (2019).
- Bond, R. M. et al. A 61-million-person experiment in social influence and political mobilization. *Nature* **489**, 295–298 (2012).
- Eagle, N. & Pentland, A. S. Reality mining: sensing complex social systems. *Personal. ubiquitous Comput.* **10**, 255–268 (2006).
- Aharony, N., Pan, W., Ip, C., Khayal, I. & Pentland, A. Social fMRI: investigating and shaping social mechanisms in the real world. *Pervasive Mob. Comput.* **7**, 643–659 (2011).
- Sapiezynski, P., Stopczynski, A., Lassen, D. D. & Lehmann, S. Interaction data from the Copenhagen Networks Study. *Sci. Data* **6**, 1–10 (2019).
- The White House, Office of the Press Secretary. Statement by the president. <https://obamawhitehouse.archives.gov/the-press-office/2013/06/07/statement-president> (2013).
- Guidelines 04/2020 on the use of location data and contact tracing tools in the context of the covid-19 outbreak. https://edpb.europa.eu/sites/edpb/files/files/file1/edpb_guidelines_20200420_contact_tracing_covid_with_annex_en.pdf (2020).
- Althuler, Y., Aharony, N., Fire, M., Elovici, Y. & Pentland, A. Incremental learning with accuracy prediction of social and individual properties from mobile-phone data. In *Proceedings of the 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust*. 969–974 (IEEE, 2012).
- Blumenstock, J., Cadamuro, G. & On, R. Predicting poverty and wealth from mobile phone metadata. *Science* **350**, 1073–1076 (2015).
- Luo, S., Morone, F., Sarraute, C., Travizano, M. & Makse, H. A. Inferring personal economic status from social network location. *Nat. Commun.* **8**, 1–7 (2017).
- Chamberlain, B. P., Humby, C. & Deisenroth, M. P. Probabilistic inference of twitter users' age based on what they follow. In Altun Y. et al. (eds) *Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2017. Lecture Notes in Computer Science*, Vol 10536 (Springer, Cham, 2017).
- Felbo, B., Sundsoy, P., Pentland, S. A., Lehmann, S. & de Montjoye, Y.-A. Modeling the temporal nature of human behavior for demographics prediction. In Altun Y. et al. (eds) *Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2017. Lecture Notes in Computer Science*, vol 10536 (Springer, Cham, 2017).
- Singh, V. K., Freeman, L., Lepri, B. & Pentland, A. S. Predicting spending behavior using socio-mobile features. In *IEEE International Conference on Social Computing* (2013).
- de Montjoye, Y.-A., Quoidbach, J., Robic, F. & Pentland, A. S. Predicting personality using novel mobile phone-based metrics. In *Greenberg A.M., Kennedy W.G., Bos N.D. (eds) Social Computing, Behavioral-Cultural*

- Modeling and Prediction. SBP 2013. Lecture Notes in Computer Science*, vol 7812 (Springer, Berlin, Heidelberg, 2013).
21. Gao, J., Zhang, Y.-C. & Zhou, T. Computational socioeconomics. *Phys. Rep.* **817**, 1–104 (2019).
 22. Humbert, M., Trubert, B. & Huguenin, K. A survey on interdependent privacy. *ACM Comput. Surv.* **52**, (2019).
 23. Opsahl, K. Electronic Frontier Foundation. Why metadata matters. <https://www.eff.org/deeplinks/2013/06/why-metadata-matters> (2019).
 24. Schoen, S. What location tracking looks like. *Electron. Front. Foundation* <https://www.eff.org/deeplinks/2011/03/what-location-tracking-looks> (2011).
 25. Greenwald, G. NSA collecting phone records of millions of Verizon customers daily. *The Guardian* <https://www.theguardian.com/world/2013/jun/06/nsa-phone-records-verizon-court-order> (2013).
 26. Stempel, J. NSA's phone spying program ruled illegal by appeals court. *Reuters* <https://www.reuters.com/article/us-usa-security-nsa-idUSKBNONS1I20150507> (2015).
 27. House of Commons and the House of Lords—Joint Committee on Human Rights. *Human Rights and the Government's Response to Covid-19: Digital Contact Tracing* <https://publications.parliament.uk/pa/jt5801/jtselect/jtrights/343/343.pdf> (2020).
 28. Sharma, T. & Bashir, M. Use of apps in the covid-19 response and the loss of privacy protection. *Nat. Med.* **26** (8), 1165–1167 (2020).
 29. Norwegian Institute of Public Health. *NIPH Stops Collection of Personal Data in Smittestopp* <https://www.fhi.no/en/news/2020/niph-stops-collection-of-personal-data-in-smittestopp/> (2020).
 30. European Parliament, Council of the European Union. *General Data Protection Regulation* <https://eur-lex.europa.eu/eli/reg/2016/679/oj> (2016).
 31. Article 29 Data Protection Working Party. *Opinion 05/2014 on Anonymisation Techniques* https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf (2014).
 32. BT. *BT.com Privacy Policy* (2018, accessed 30 March 2020) https://img01.products.bt.co.uk/content/dam/bt/storefront/pdfs/BT.comcurrentprivacypolicy_18052018.pdf.
 33. O2. *Our Privacy Policy* (2020, accessed 16 December 2020) <https://www.o2.co.uk/termsandconditions/privacy-policy>.
 34. European Data Protection Board. *Guidelines 4/2019 on Article 25—Data Protection by Design and by Default* (2020) https://edpb.europa.eu/sites/edpb/files/consultation/edpb_guidelines_201904_dataprotection_by_design_and_by_default.pdf.
 35. Sweeney, L. Weaving technology and policy together to maintain confidentiality. *J. Law, Med. Ethics* **25**, 98–110 (1997).
 36. Narayanan, A. & Shmatikov, V. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy*. 111–125 (IEEE, 2008).
 37. de Montjoye, Y.-A., Hidalgo, C. A., Verleysen, M. & Blondel, V. D. Unique in the crowd: the privacy bounds of human mobility. *Sci. Rep.* **3**, 1376 (2013).
 38. Riederer, C., Kim, Y., Chaintreau, A., Korula, N. & Lattanzi, S. Linking users across domains with location data: theory and validation. In *Proc. 25th International Conference on World Wide Web*, 707–719 (International World Wide Web Conferences Steering Committee, 2016).
 39. Backstrom, L., Dwork, C. & Kleinberg, J. Wherefore art thou R3579X? Anonymized social networks, hidden patterns, and structural steganography. In *Proc. 16th International Conference on World Wide Web*, 181–190 (ACM, 2007).
 40. Hay, M., Miklau, G., Jensen, D., Weis, P. & Srivastava, S. Anonymizing Social Networks. *Computer Science Department Faculty Publication Series* 180 (2007).
 41. Zhou, B. & Pei, J. Preserving privacy in social networks against neighborhood attacks. In *2008 IEEE 24th International Conference on Data Engineering*, 506–515 (2008).
 42. Zou, L., Chen, L. & Özsu, M. T. K-automorphism: a general framework for privacy preserving network publication. *Proc. VLDB Endow.* **2**, 946–957 (2009).
 43. Cheng, J., Fu, A. W.-c. & Liu, J. K-isomorphism: privacy preserving network publication against structural attacks. In *Proc. 2010 ACM SIGMOD International Conference on Management of Data*, 459–470 (ACM, 2010).
 44. Wang, G., Liu, Q., Li, F., Yang, S. & Wu, J. Outsourcing privacy-preserving social networks to a cloud. In *2013 Proceedings IEEE INFOCOM*, 2886–2894 (IEEE, 2013).
 45. Liu, Q., Wang, G., Li, F., Yang, S. & Wu, J. Preserving privacy with probabilistic indistinguishability in weighted social networks. *IEEE Trans. Parallel Distrib. Syst.* **28**, 1417–1429 (2016).
 46. Gao, J., Ping, Q. & Wang, J. Resisting re-identification mining on social graph data. *World Wide Web* **21**, 1759–1771 (2018).
 47. Narayanan, A. & Shmatikov, V. De-anonymizing social networks. In *2009 30th IEEE Symposium on Security and Privacy*, 173–187 (IEEE, 2009).
 48. Nilizadeh, S., Kapadia, A. & Ahn, Y.-Y. Community-enhanced de-anonymization of online social networks. In *Proc. 2014 ACM SIGSAC Conference on Computer and Communications Security*, 537–548 (ACM, 2014).
 49. Sharad, K. & Danezis, G. An automated social graph de-anonymization technique. In *Proc. 13th Workshop on Privacy in the Electronic Society*, 47–58 (ACM, 2014).
 50. Ji, S., Li, W., Gong, N. Z., Mittal, P. & Beyah, R. A. On your social network de-anonymizability: quantification and large scale evaluation with seed knowledge. In *NDSS* (Internet Society, 2015).
 51. Gulyás, G. G., Simon, B. & Imre, S. An efficient and robust social network de-anonymization attack. In *Proc. 2016 ACM on Workshop on Privacy in the Electronic Society*, 1–11 (ACM, 2016).
 52. Sharad, K. Change of guard: the next generation of social graph de-anonymization attacks. In *Proc. 2016 ACM Workshop on Artificial Intelligence and Security*, 105–116 (ACM, 2016).
 53. Shao, Y., Liu, J., Shi, S., Zhang, Y. & Cui, B. Fast de-anonymization of social networks with structural information. *Data Sci. Eng.* **4**, 76–92 (2019).
 54. Srivatsa, M. & Hicks, M. Deanonymizing mobility traces: Using social network as a side-channel. In *Proc. 2012 ACM Conference on Computer and Communications Security*, 628–637 (ACM, 2012).
 55. Pedarsani, P., Figueiredo, D. R. & Grossglauser, M. A bayesian method for matching two similar graphs without seeds. In *2013 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 1598–1607 (IEEE, 2013).
 56. Yartseva, L. & Grossglauser, M. On the performance of percolation graph matching. In *Proc. first ACM Conference on Online Social Networks*, 119–130 (ACM, 2013).
 57. Korula, N. & Lattanzi, S. An efficient reconciliation algorithm for social networks. *Proc. VLDB Endow.* **7**, 377–388 (2014).
 58. Fabiana, C., Garetto, M. & Leonardi, E. De-anonymizing scale-free social networks by percolation graph matching. In *2015 IEEE Conference on Computer Communications (INFOCOM)*, 1571–1579 (IEEE, 2015).
 59. Man, T., Shen, H., Liu, S., Jin, X. & Cheng, X. Predict anchor links across social networks via an embedding approach. In *IJCAI*, Vol. 16, 1823–1829 (AAAI Press, 2016).
 60. Zhang, W., Shu, K., Liu, H. & Wang, Y. Graph neural networks for user identity linkage. Preprint at *arXiv* <https://arxiv.org/abs/1903.02174> (2019).
 61. Bronstein, M. M., Bruna, J., LeCun, Y., Szlam, A. & Vandergheynst, P. Geometric deep learning: going beyond euclidean data. *IEEE Signal Process. Mag.* **34**, 18–42 (2017).
 62. Veličković, P. et al. Graph attention networks. In *Sixth International Conference on Learning Representations (ICLR)*, (2018).
 63. Bruna, J., Zaremba, W., Szlam, A. & LeCun, Y. Spectral networks and deep locally connected networks on graphs. In *Second International Conference on Learning Representations, 2014* (ICLR, 2014).
 64. Defferrard, M., Bresson, X. & Vandergheynst, P. Convolutional neural networks on graphs with fast localized spectral filtering. In *Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R. (eds) Advances in Neural Information Processing Systems*, 3844–3852 (NIPS, 2016).
 65. Kipf, T. N. & Welling, M. Semi-supervised classification with graph convolutional networks. In *Fifth International Conference on Learning Representations* (ICLR, 2017).
 66. Hamilton, W., Ying, Z. & Leskovec, J. Inductive representation learning on large graphs. In *Guyon et al. (eds) Advances in Neural Information Processing Systems*, 1024–1034 (NIPS, 2017).
 67. Zitnik, M., Agrawal, M. & Leskovec, J. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics* **34**, i457–i466 (2018).
 68. Veselkov, K. et al. Hyperfoods: Machine intelligent mapping of cancer-beating molecules in foods. *Sci. Rep.* **9**, 1–12 (2019).
 69. Gonzales, G., Gong, S., Laponogov, I., Veselkov, K. & Bronstein, M. Graph attentional autoencoder for anticancer hyperfood prediction. *NeurIPS 2019 Workshop on Graph Representation Learning* (2019).
 70. Gligorijević, V. et al. Structure-based protein function prediction using graph convolutional networks. *Nat. Commun.* **12**, 1–14 (2021).
 71. Gainza, P. et al. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nat. Methods* **17**, 184–192 (2020).
 72. Spier, N. et al. Classification of polar maps from cardiac perfusion imaging with graph-convolutional neural networks. *Sci. Rep.* **9**, 1–8 (2019).
 73. de Montjoye, Y.-A., Rocher, L. & Pentland, A. S. bandicoot: a python toolbox for mobile phone metadata. *J. Mach. Learn. Res.* **17**, 6100–6104 (2016).
 74. Mnih, V. et al. Recurrent models of visual attention. *Adv. Neural Inf. Process. Syst.* **27**, 2204–2212 (2014).
 75. Bahdanau, D., Cho, K. & Bengio, Y. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations* (ICLR, 2015).
 76. Schroff, F., Kalenichenko, D. & Philbin, J. Facenet: A unified embedding for face recognition and clustering. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 815–823 (IEEE, 2015).
 77. Blondel, V. D. et al. Data for development: the d4d challenge on mobile phone data. Preprint at *arXiv* <https://arxiv.org/abs/1210.0137> (2012).

78. NHSX. *Documentation Relating to the Beta of the NHS COVID-19 app* <https://github.com/nhsx/COVID-19-app-Documentation-BETA/> (2020).
79. INRIA. *Publication of the ROBERT (ROBust and privacy-presERving proximity Tracing) Protocol* <https://www.inria.fr/en/publication-robot-protocol> (2020).
80. PACT. *Pact: Private Automated Contact Tracing* <https://pact.mit.edu/> (2020).
81. Troncoso, C. et al. *Decentralized Privacy-preserving Proximity Tracing* <https://github.com/DP-3T/documents/blob/master/DP3T%20White%20Paper.pdf> (2020).
82. Apple & Google. *Privacy-preserving Contact Tracing* <https://www.apple.com/covid19/contacttracing> (2020).
83. Canetti, R., Trachtenberg, A. & Varia, M. Anonymous collocation discovery: harnessing privacy to tame the coronavirus. Preprint at *arXiv* <https://arxiv.org/abs/2003.13670> (2020).
84. Castellucia, C. et al. DESIRE: a third way for a european exposure notification system leveraging the best of centralized and decentralized systems. <https://hal.inria.fr/hal-02570382/document> (2020).
85. Verhelst, K. et al. *Proposition de loi relative à l'utilisation d'applications numériques de traçage de contacts par mesure de prévention contre la propagation du coronavirus covid-19 parmi la population* <https://www.lachambre.be/FLWB/PDF/55/1251/55K1251001.pdf> (2020).
86. President's Council of Advisors on Science and Technology. *Big Data and Privacy: a Technological Perspective* https://bigdatawg.nist.gov/pdf/pcast_big_data_and_privacy_-_may_2014.pdf (2014).
87. de Montjoye, Y.-A. et al. On the privacy-conscious use of mobile phone data. *Sci. Data* **5**, 1–6 (2018).
88. Thielman, S. Surveillance reform explainer: can the FBI still listen to my phone calls? *The Guardian* <https://www.theguardian.com/world/2015/jun/03/surveillance-reform-freedom-act-explainer-fbi-phone-calls-privacy> (2015).
89. Hermans, A., Beyer, L. & Leibe, B. In defense of the triplet loss for person re-identification. Preprint at *arXiv* <https://arxiv.org/abs/1703.07737> (2017).

Acknowledgements

A.-M.C. is the recipient of a Ph.D. scholarship from Imperial College London's Department of Computing. X.D. gratefully acknowledges support from the Oxford-Man Institute of Quantitative Finance. The authors would like to thank the Computational Privacy Group and, in particular, Luc Rocher, Florimond Houssiau, Andrea Gadotti, and Shubham Jain for their valuable feedback and helpful discussions.

Author contributions

Y.-A.d.M. and A.-M.C. conceived the attack. A.-M.C. analyzed the data. A.-M.C., F.M., X.D., and M.B. designed the method. A.-M.C., F.M., S.M., X.D., M.B., and Y.-A.d.M. designed the experiments. A.-M.C. and S.M. performed the experiments. A.-M.C., M.B., and Y.-A.d.M. wrote the paper with input from all the authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-021-27714-6>.

Correspondence and requests for materials should be addressed to Yves-Alexandre de Montjoye.

Peer review information *Nature Communications* thanks Kévin Huguenin, Petar Veličković, Dashun Wang and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022