Using Location as a Signal to Affirm Identity Online

Roberto Colecchia, Rafael Gouveia, Filipe Martins, Hector Pinheiro, Matheus Leite, Raiza Oliveira
Incognia
2479 East Bayshore Road, Suite 150
Palo Alto, CA 94303, USA
(650) 463-9280

Abstract - In this paper we propose a method to estimate online user Identity Uniqueness based on the user location history. We show that a user's location history can be utilized as a signal to affirm identity online. The accuracy of Identity Uniqueness based on location is shown to be in the range of 1:1.1M to 1.:17M, which is higher than the FAR (False Acceptance Rate) of leading physical biometric factors in commercial mobile applications (1:50K to 1:1M)

Keywords: user location history, identity uniqueness, identity affirmation, mobile location

I. Introduction

Identity proofing is an essential component of digital security and establishes that a user is who they claim to be. According to Gartner [1], identity proofing refers to the combination of activities during an online interaction that brings a real-world identity claim and aims to assure that:

- The real-world identity exists.
- The individual claiming the identity is the true owner of that identity and is genuinely present during the process.

NIST (National Institute of Standards and Technology) similarly defines identity proofing as the process by which organizations collect sufficient information to validate, and verify the identity of a person [2]. As part of identity proofing, Gartner further defines identity affirmation [1] as a set of capabilities that:

- Assess whether a real-world identity exists for an online identity claim.
- Can provide supporting risk or trust signals to an online identity claim.

Identity affirmation plays an important role when users register with online systems and when users are trying to authenticate remotely. It's important to have scalable & reliable technologies to verify a user's online identity to improve fraud prevention but at the same time minimize user friction and respect privacy and applicable data protection legislation.

The current identity affirmation approaches described in [1] include use of database references, and information on the user's device, phone number, email and location. Widespread theft of Personally Identifiable Information (PII) via data breaches and phishing has eroded the reliability of methods that rely on static data. On the other hand, if we consider location behavior i.e. the user's information based on the history of locations detected on the user's mobile phone - it is much more robust against fraud

since it cannot be phished and is dynamic and constantly updating. Location data provides improved strength against fraud while at the same time ensuring a lower friction (or zero-friction) experience for the user since it can be automatically derived from the user's smartphone.

In this paper, we will describe how it is possible to

- Use location to establish the uniqueness of an identity;
- Affirm the online identity of an individual using his/her location history.

In particular, we will analyze the accuracy provided by user location history in defining the uniqueness of the identity and compare it with the FAR (False Acceptance Rate) provided by physical biometric factors. We will show how the accuracy of location-based Identity Uniqueness is comparable to, and even higher than, the FAR of physical biometric factors for consumer mobile applications.

II. Prior Work

In this paper, we are exploring the use of location data as a signal for identity affirmation. Some research has been done in the past about the possibility to determine *Identity Uniqueness* based on location information: this prior research has studied how location history can be used to identify a subject or an individual.

Two research study examples that inspired this paper are [3] and [4]. In [3], researchers have considered a static database of users' locations based on cell phone cells/tower data collected in Europe over 15 months. Cell Tower Area resolution size may vary, but in general, they are of about 10 miles (~16Km) radius for the 3G networks used for this analysis. The researchers have concluded that, when considering traces with the location resolution of *Cell Tower Level*, 11 historic location points are required to uniquely identify an individual (within a set of 1.5M mobile users).

In [4], researchers have used static U.S. Census data from 2004 to determine if individuals could be identified based only on their home and work locations - registered by the U.S. Census, using a location resolution of *Census Tract*. Census Tract size may vary since they depend on the population density. However, they are roughly similar to zip codes (~10-90 sq miles / ~25-230Km^2). The researchers have concluded that with the knowledge of two locations (home and work) at this coarse location resolution level, about 5% of the U.S. employed population (which totals ~150M) is uniquely identifiable, and the majority (50%+) of the U.S. employed population shares these location traits with 10 other people or less. The latest U.S. Census public databases provide this individual home/work information also at a *City Block* level - called a *Census Block* - and allow for a more granular level of location analysis. We will further explore this later in this paper (see chapter V.).

III. Parameters affecting accuracy for Location-based Identity Uniqueness

In general, the prior works mentioned above agree on the fact that a subject can be

uniquely identified by studying data about the location he/she visited during a given temporal interval. However, in the literature, we could not find any prior study that tried to establish or measure the accuracy of Identity Uniqueness. In this paper, we will identify two parameters that can help determine the identity of the user based on his/her location history.

1) Number of "Location Points" considered for each user: in our study, a "location" is a place that a user visits frequently and where he/she usually spends some time. So, for example, "Home" can be the user's location #1, "Work" can be the user;s location #2, etc. The more locations points are available, the higher is the accuracy for the Identity Uniqueness, enabling the identification of one user versus the other more easily.

2) Resolution of the "Location Unit" considered: the previous studies mentioned above were based on somewhat coarse location unit resolution: Cell Tower Coverage Size (~10 Miles, ~16 Km radius) and Census Tract Area Size (~10-90 sq miles, ~25-230 Km^2). Incognia leverages GPS, WiFi and other signals based on the user's smartphone sensors, so the location unit resolution is much higher. In this study, we have considered Incognia data based on GPS precision level, to evaluate the accuracy (50m radius, as described in VII). The higher the resolution of the location unit considered, the higher is the accuracy for the Identity Uniqueness, enabling easier identification of one user from another.

To define the resolution of the "Location Unit" in this study, we have used the geohash public domain geocode system described in [5]. As a summary, in Table 1 are the key sizes for location unit:

Table 1
Geohash geocode system resolution

Geohash Level	Location Unit Resolution	Approximate Real World Correspondence
4	39.1km x 19.5km	Census Tract / Zip Code
5	4.9km x 4.9km	Cell Tower Coverage
6	1.2km x 609.4m	
7	152.9m x 152.4m	Census Block / City Block
8	38.2m x 19m	Single Building Size
9	4.8m x 4.8m	Single Room Size

In this study, we will define as $\mathbf{k_{n,m}}$ the accuracy for the Identity Uniqueness obtained with **n** location points and geohash level = **m.**

One of the key commonalities between studies [3] and [4] is that both try to measure Identity Uniqueness based on static databases. In [3], location points = 11 and the location unit resolution is comparable with geohash level 5 ($\mathbf{k_{11,5}}$ = 1:1.5M), while in [4], location points = 2 and the location unit resolution is comparable with geohash level 4 ($\mathbf{k_{2,4}}$ = varies between 1:150M and 1:150K).

In real-world authentication applications, the location database is dynamic and keeps changing with every new data point

received. For example, people may move or travel, and the authentication needs to adapt to it. So using static databases can provide a "picture" of the Identity Uniqueness at a given moment in time, but using dynamic location databases (such as Incognia) can help to leverage Identity Uniqueness in real-time for authentication applications.

In this paper, we will investigate the accuracy of Identity Uniqueness based on location determined by the following configurations:

- Number of location points = 2, 3, 4, 5
- Location unit resolution = geohash level 7 (city block resolution, as highlighted in Table 2)

Therefore, determining $\mathbf{k_{2,7}}$, $\mathbf{k_{3,7}}$, $\mathbf{k_{4,7}}$, and $\mathbf{k_{5,7}}$

Table 2
Accuracy of Identity Uniqueness based on number of location points

Number of location points	Geohash level 4 (Zip Code)	Geohash level 5 (Cell Phone Tower)	Geohash level 7 (City Block)
2	$\mathbf{k_{2,4}}$ Study *(2)* between **1:150M** and **1:150K** for different sets of US employed population		$\mathbf{k_{2,7}}$
3			$\mathbf{k_{3,7}}$
4			$\mathbf{k_{4,7}}$
5			$\mathbf{k_{5,7}}$
11		$\mathbf{k_{11,5}}$ Study *(1)* **1:1.5M**	

To do this, we will use the U.S. Census static data and data gathered by Incognia. Study [4] uses home and work location: this is in conformity with Incognia's approach, which consists of gathering location information for routine and usual user behavior. In section V, we will be using U.S. Census data with greater granularity than that which has been used in [4].

IV. Anonymity Sets and Identity Uniqueness

As an example, in the case of **n** = 2 (number of location points), using user location data, it is possible to determine how many people live in a specific location unit "A" (1st location) and work in a specific location unit "B" (2nd location). The subject's location trace is the only one with the home/workplace pair (A, B). To describe this, we will use the model proposed in [4] and [6] defining *Anonymity Sets*. The set of all people associated with the pair (A, B) is called the *anonymity set* of the pair. The larger the anonymity set, the larger the crowd one is indistinguishable from, and consequently, the more difficult it is to identify him/her. Enlarging the size of regions A and/or B (by changing the location unit resolution, for example, geohash levels from 5 to 4) increases the size of the anonymity set and thus the difficulty of subject identification. If the size of anonymity set = 1, then the subject is fully identified by the home/workplace pair (A, B). If the size of the anonymity set = 3, three individuals share the same home/workplace pair (A, B), and so on.

Fig. 1 and Fig. 2 illustrate how the size of the anonymity set on the same population

changes with the change of the location resolution. In Fig. 1 the location unit, represented by the gray rectangles, is a geohash of level 5, which is approximately a *Zip Code*. In Fig. 2 the location unit, shown by the gray rectangles, is a geohash of level 7, which is approximately a *City Block*. H1, H2, H3 and W1, W2, W3 indicate the home and work locations for persons 1, 2 and 3.

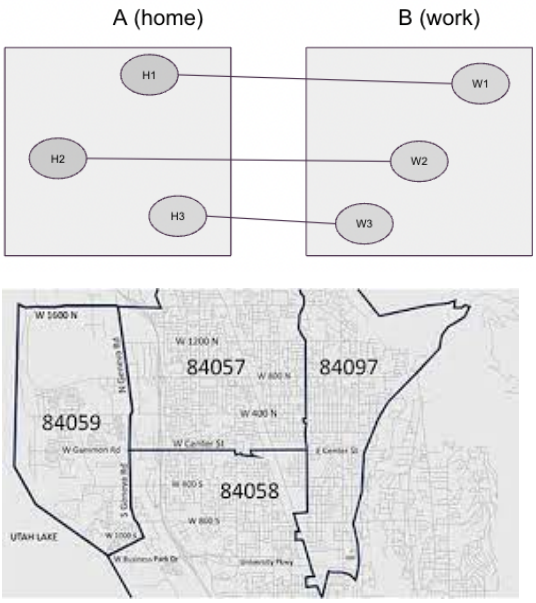

Fig. 1: Low location resolution anonymity set (i.e. Zip Code Level)

n = 2 (2 location points: home/work)
m = 5 (geohash level 5 location resolution: approximation of *Zip Code*)
Size of anonymity set = 3 (3 individuals share the pair (A,B) for home/work)

Fig. 2: High location resolution anonymity set (i.e. City Block Level)

n = 2 (2 location points: home/work)
m = 7 (geohash level 7 location resolution: approximation of a *City Block*)
Size of anonymity set = 1 (1 individual is fully identified by the pair (A,B) for home/work)

V. Determining the location-based Identity Uniqueness accuracy using the U.S. Census LEHD (Longitudinal Employer-Household Dynamics) Database

In order to study the accuracy $\mathbf{k_{2,7}}$ of location-based Identity Uniqueness (defined in Table 2), we have used static data from the U.S. Census reported in 2019, the latest available dataset to date. The U.S. Census Bureau has publicly posted the available data about the U.S. population, and it can be downloaded from [7].

In the LEHD (Longitudinal Employer Household Dynamics) database is collected the anonymized information of the U.S. working population, in terms of each individual's home and work addresses. Among other uses, the U.S. government

employs this data to analyze commuter patterns to see how much commuting is going on in every area and eventually use it as supporting information to build new traffic infrastructure e.g. roads, highways, etc.. The information is available with location unit resolution of *Census Tract* (approximately similar to *Zip Code* or geohash level 4) and with location unit resolution of *Census Block* (approximately similar to a *City Block* or geohash level 7).

In [4], there is an analysis of the sizes of anonymity sets based on *Census Tracts* using the LEHD database (based on 2004 U.S. Census data). The anonymity set size is much smaller when both home/work locations are considered than if only one of these (home or work) is considered as shown in Fig. 3.

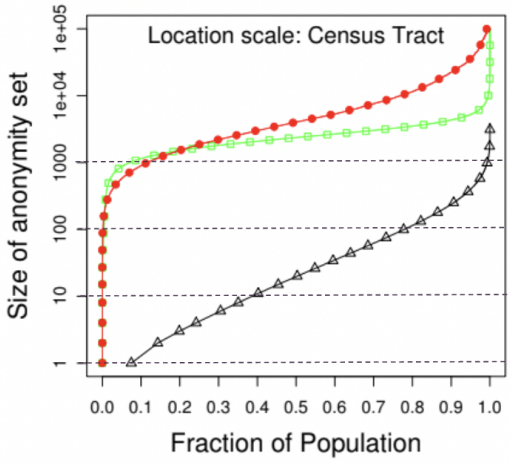

Fig. 3: Size of the anonymity set under the disclosure of work location (red circles), home location (green squares) or both (black triangles). Location granularity is the *Census Tract*.

So, suppose we try to extrapolate $\mathbf{k}_{2,4}$ (**n** = 2 location points, **m** = geohash level 4) based on this graph. In this case, 100% of the U.S. working population (which totals about 150M people) belongs to anonymity sets with size < 1000, 50% of this population belongs to anonymity sets with size < 10, and 5% of this population is uniquely identified by home/work pair. So we could estimate that $\mathbf{k}_{2,4}$ varies between 1:150M and 1:150K.

In this paper, we want to explore the accuracy at higher resolutions for location units, so we have considered the LEHD database at the *Census Block* level (approximately the same size as geohash level 7) rather than at *Census Tract* level (approximately corresponding to geohash level 4) as done in [4]. So, we considered the *Census Block* as location unit resolution and downloaded the corresponding data from California (~17.1M employed records available) from the LEHD database.

Summarized in Table 3 are the results from the top 10 anonymity sets from the California LEHD 2019 database, for 2 location points (n = 2) and Census Blocks location unit size resolution (m = geohash 7). In this case, we see that:

- About 87% of the employed individuals (~14.9M) are uniquely identified by their home block and work block pair (anonymity set = 1).
- 7% of the employed individuals (~652K) have another person with whom they share the same home/work block pair (anonymity set = 2)
- 2% of the employed individuals (~128K) share the same home/work block pair with the other two individuals (anonymity set = 3).

Table 3
Anonymity sets and collisions for California LEHD 2019 dataset (U.S. Census)

Size of anonymity set	Frequency of home/work Census block pairs	Total People	Percentage	Collisions
1	14,947,672	14,947,672	87.01%	0
2	652,546	1,305,092	7.60%	1305092
3	128,457	385,371	2.24%	1156113
4	44,061	176,244	1.03%	1057464
5	19,521	97,605	0.57%	976050
6	10,041	60,246	0.35%	903690
7	5,886	41,202	0.24%	865242
8	3,624	28,992	0.17%	811776
9	2,433	21,897	0.13%	788292
10	1,669	16,690	0.10%	751050

The graph in Fig. 4 summarizes the distribution of anonymity set sizes for 2 location points (**n** = 2) and Census Blocks location unit size resolution (**m**= ~geohash level 7):

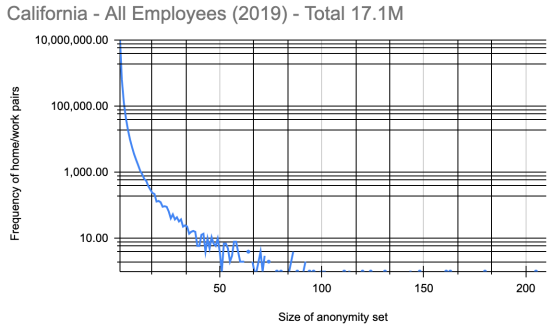

Fig. 4: Distribution of Anonymity set sizes for 2 location points

VI. Collision between identities and accuracy of Identity Uniqueness

To measure the accuracy of Identity Uniqueness more precisely, we have also defined the term "collision of identities" as the number of distinct identities mapped into a single subject.

So, to define a measure for the accuracy of Identity Uniqueness, we need to determine how to assess **c** as the probability of collisions between identities for each anonymity set - given a location unit resolution and a given number of location points. We propose the following formula:

$$(\alpha)\ c_a = \frac{s_a \times (s_a - 1)}{2} \times t_a$$

where
a = anonymity set number
$\mathbf{s_a}$ = size of anonymity set **a**
$\mathbf{c_a}$ = number of identity collisions for anonymity set **a**
$\mathbf{f_a}$= total people in anonymity set **a**

A collision is generated for all pairs (A, B) in the anonymity set where A ≠ B

Given the collisions for each one of the anonymity sets present in the distribution, we can calculate the accuracy **k**, based on the proposed formula below:

$$(\beta)\ k_{n,m} = 1 - (\Sigma\ c_i\ /\ (\frac{t \times (t-1)}{2}))$$

where
k = accuracy
n = number of location points
m = location unit resolution (geohash level)
$\mathbf{c_i}$ = number of identity collisions for each anonymity set **i**
t = total number of subjects

In (β) the term t*(t-1) comes from the fact that in this analysis we are comparing everyone with everyone else. Other kinds of criteria could be chosen, like for example only comparing people living in the same city or region. This should reduce the denominator and change the proposed accuracy: this could be further explored in further research or in a next paper. However, we consider this current assumption to be fair in comparison to the studies presented as reference, since in both cases the uniqueness values are regarding the whole population. Based on these definitions, we have calculated:

$\mathbf{k}_{2,7} = 0.9999995757 = 1 : 2.3M$

VII. Accuracy calculations based on Incognia Location Data-set

To measure the accuracy for different numbers of location points and different location unit resolutions, we have used the Incognia Location Data-set.

The Incognia SDK is installed on more than 200M mobile devices (iOS, Android). For each device, according to user's permissions, Incognia collects information about the device's "frequent" or "trusted" locations: although Incognia does not know what these locations refer to ("home", "work", "parents house", "school", "library", etc), these are locations where the device spends most of its time. In this paper we will not discuss the details regarding the methods for selecting such places from a base of raw location points from each device.

Incognia detected the device location, based on the GPS available on the user cell phones. The resolution used is described by a circle with a radius of 50m (100m diameter) which is similar to geohash level 7 (152.9m x 152.4m), and it is approximated using hexagons (as done in H3, the hierarchical spatial index by Uber Technologies Open Source location descriptor in [8]) like shown on the image below. This enables detecting locations at the level of *Census Block*, similar to what is provided by the latest U.S. Census database.

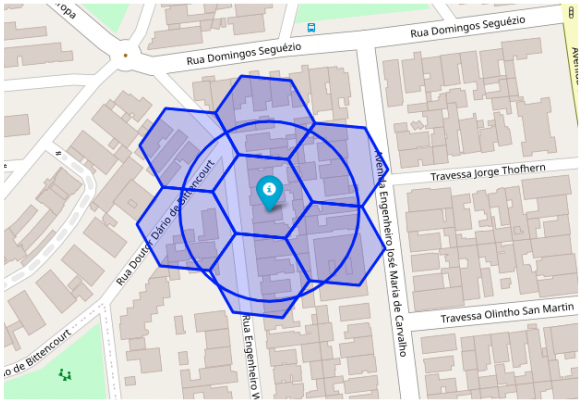

Fig. 5: Resolution used with Incognia data: 100m diameter, similar to geohash level 7, and Census Block

Using the Incognia Location data-set, on 10M users, with 2-3-4-5 location points, with a location resolution of geohash level 7 and with the collision and accuracy formulas in (α) (β), we find these values:

$\mathbf{k}_{2,7} = 1 : 1.1M$
$\mathbf{k}_{3,7} = 1 : 1.9M$
$\mathbf{k}_{4,7} = 1 : 5M$
$\mathbf{k}_{5,7} = 1 : 17M$

Table 4
Accuracy of Identity Uniqueness based on Incognia data-set with resolution diameter 100m, ~Geohash level = 7, ~ Census Block

Location points = 2	Location points = 3	Location points = 4	Location points = 5
1 : 1.1M ($\mathbf{k}_{2,7}$) 1 : 2.3M ($\mathbf{k}_{2,7}$)*	1 : 1.9M ($\mathbf{k}_{3,7}$)	1 : 5M ($\mathbf{k}_{4,7}$)	1 : 17M ($\mathbf{k}_{5,7}$)

* = based on California LEHD 2019 dataset (U.S. Census)

These results are also shown in Fig. 6.

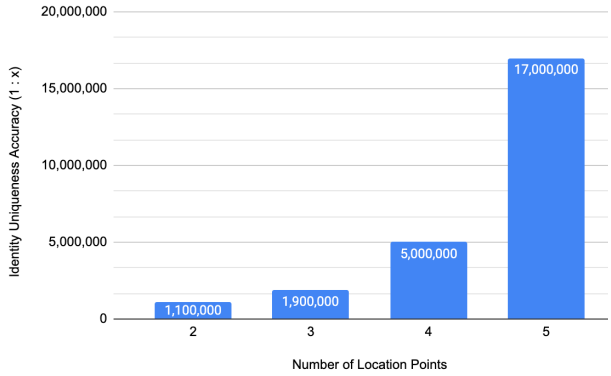

Fig. 6: Accuracy of location for Identity Uniqueness based on number of location points

The result obtained in estimating the accuracy may depend on several factors:

- Data volume per each user available.
- Number of users in analysis (population density).
- How we define a "frequent location"
- Criteria to create comparisons between users or groups of users. In our case we have compared everyone with everyone else looking for uniqueness.
- Criteria to consider two sets of geographical coordinates as the same location (location matching algorithms).

These factors may partially explain the difference between $\mathbf{k}_{2,7}$ calculated using static U.S. Census data, versus $\mathbf{k}_{2,7}$ calculated with Incognia dynamic database. We also need to consider that the *Census Block* does not have a fixed dimension but its size is variable and depends on population density.

VIII. Comparison with the FAR of leading consumer Biometric Technologies

Fig. 7 shows that with a suitable number of locations and with a high enough location unit resolution, the accuracy of Location for Identity Uniqueness is comparable or higher than the False Acceptance Rate (FAR) for physical biometric technologies. In fact the accuracy shown for 2,3,4,5 or more locations points and geohash level 7 (City Block Size) varies between $\mathbf{k}_{2,7}$ = 1 : 1.1M and $\mathbf{k}_{5,7}$ = 1 : 17M.

For the leading biometric technologies in consumer products - Apple and Microsoft, as reported in [9], [10] and [11] - the FAR for fingerprint recognition (Touch ID and Windows Hello) is in the range of 1: 50K while for face recognition (Face ID and Windows Hello) is in the range of 1:1M.
By comparison, the odds of guessing a typical 4-digit passcode are 1 in 10,000.

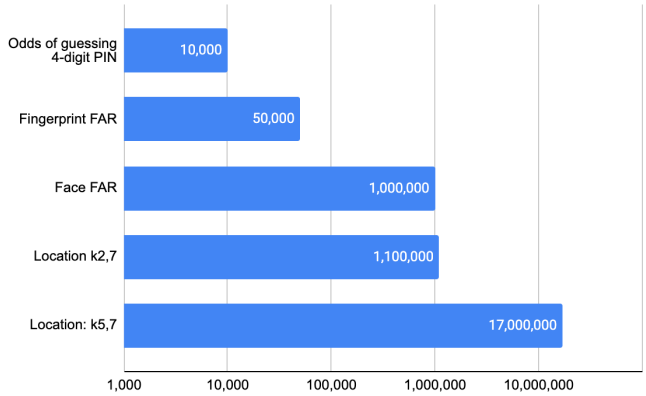

Fig. 7: Location Identity Uniqueness Accuracy ($\mathbf{k}_{2,7}$ and $\mathbf{k}_{5,7}$ with Incognia data) & FAR for Leading Biometric Technologies

IX. Conclusion

From the data collected, the reader can see that the location Identity Uniqueness accuracy

- increases with the number of the available location points
- increases with the resolution of the location unit.

Given these results, location technology with an appropriate number of location points and an appropriate resolution, can be considered a valid and robust means to affirm identity for online consumer applications. The accuracy that can be achieved for the Identity Uniqueness is higher than FAR (False Acceptance Rate) of physical biometrics technologies in mobile consumer applications.

Furthermore, when determining the location with the use of WiFi in addition to GPS, it is possible to explore the accuracy of location uniqueness at geohash level 8 (38.2m x 19.4 m - a single building size) and high accuracy results (such as for example 1: 5M and 1:17M) are expected with even fewer trusted locations points. We plan to explore the accuracy of Identity Uniqueness with WiFi resolution and geohash level 8 in a following paper.

X. Incognia Privacy Policy

Incognia promotes the mission of developing high performance technology without storage or access to data that can directly identify users, since Incognia does not collect unique static identifiers from mobile devices, associated accounts or civil identification data. Incognia follows the fundamental principles of Privacy by Design, has SOC 2 Type II Certification and complies with LGPD (Brazilian Privacy Law), CCPA (California Privacy Law), GDPR (European Union Privacy Regulation). The Incognia Solution Privacy Policy [12] is available on its Website with clear, accessible and transparent information on the processing of personal data.

References

[1] Akif Khan: Market Guide for Identity Proofing and Affirmation (Gartner March 2022). https://www.incognia.com/resources/gartner-market-guide-for-identity-proofing-and-affirmation?hsLang=en

[2] Digital Identity Guidelines (NIST Special Publication 800-63-3) https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-63-3.pdf

[3] Yves-Alexandre de Montjoye, Cesar A. Hidalgo, Michel Verleysen & Vincent D. Blondel: Unique in the Crowd: The privacy bounds of human mobility (Scientific Reports 3/2013)

[4] Philippe Golle and Kurt Partridge: On the Anonymity of Home/Work Location Pairs (Palo Alto Research Center, 2009) https://crypto.stanford.edu/~pgolle/papers/commute.pdf

[5] Description of geohash for geographic location encoding (Wikipedia) https://en.wikipedia.org/wiki/Geohash

[6] L. Sweeney. K-anonymity: a Model for Protecting Privacy.(International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 557-570). https://epic.org/wp-content/uploads/privacy/reidentification/Sweeney_Article.pdf

[7] U.S. Census LEHD (Longitudinal Employer-Household Dynamics) Database: https://lehd.ces.census.gov/data/

[8] Uber Technologies H3 documentation https://h3geo.org/docs/

[9] Apple TouchID security technology https://support.apple.com/en-us/HT204587#:~:text=The%20probability%20of%20this%20happening,passcode%20are%201%20in%2010%2C000.

[10] About Face ID advanced technology https://support.apple.com/en-us/HT208108#:~:text=The%20probability%20that%20a%20random,you're%20wearing%20a%20mask.

[11] Windows Hello biometric requirements https://docs.microsoft.com/en-us/windows-hardware/design/device-experiences/windows-hello-biometric-requirements

[12] Incognia Mobile Fraud Solution - Privacy Policy https://www.incognia.com/policies/incognia-policy