Journal Article

No More Privacy Any More?

Grout, V

Recommended citation:

Grout, V (2018) 'No More Privacy Any More?', *Information*, vol 10, issue 1. Available online 09 January 2019. doi: 10.3390/info10010019 *information*

MDPI

Editorial

No More Privacy Any More?

Vic Grout

Department of Computing, Wrexham Glyndŵr University, Wrexham LLI1 2AW, UK; v.grout@glyndwr.ac.uk;
Tel.: +44-1978-293-203

Received: 4 January 2019; Accepted: 4 January 2019; Published: 9 January 2019

Abstract: The embodiment of the potential loss of privacy through a combination of artificial intelligence algorithms, big data analytics and Internet of Things technology might be something as simple, yet potentially terrifying, as an integrated app capable of recognising anyone, anytime, anywhere: effectively a global 'Shazam for People'; but one additionally capable of returning extremely personal material about the individual. How credible is such a system? How many years away? And what might stop it?

Keywords: privacy; identity theft; big data; artificial intelligence; internet of things; de-anonymization; personal identification; personal recognition; biometrics; data apps; personal data; identity voyeurism

1. Introduction: A Future Scenario?

Imagine it is 2025 or thereabouts. You meet someone at an international conference. Even before they have started to introduce themselves, your *Internet of Things* (*IoT*) connected augmented reality glasses (or something like that: it does not matter) have told you everything you needed to know . . . and a lot more you did not.

> *Jerry Gonzales. Born (02/11/1970): Glasgow, UK, dual (plus USA) citizenship; 49 years old. Married 12/12/1994 (Ellen Gonzales, nee Schwartz), divorced 08/06/2003; two daughters (Kate: 23, Sarah: 17); one son (David: 20). Previous employment: Microsoft, IBM, University of Pwllheli; current: unemployed. Health: smoker, heavy drinker, recurrent lung problems, diabetic, depression. Homeowner (previous); now public housing. Credit rating: poor (bankruptcy 10/10/2007); Insurance risk: high. Politics: Republican. etc., . . . , Sport: supports Boston Red Sox and Manchester United FC. . . . , Pornography: prefers straight but with mild abuse . . . , etc., etc.*

And that is the simple basis of this paper, along with the overlapping questions that naturally follow:

- How *likely* (futurology) is this to happen?
- What is necessary (technology) to *allow* it? And how *long* might it take?
- What can be done (legally, politically, morally, etc.) to *stop* it?

However, to begin this discussion, we consider a comparable, essentially parallel, application of technology: one that already exists, not merely legally but almost universally considered a positive use of mobile devices and the Internet.

2. A Theoretical Foundation: Shazam for People?

The music recognition system, *Shazam* [1], runs as an app on most mobile phones and tablets. Using the device's microphone, Shazam 'listens' to any (well, nearly any) piece of music for a few seconds, identifies it and informs the curious user. (It perhaps also offers the opportunity to purchase and download the track, which is not irrelevant to the discussion that follows.) How does it do this?

An indication of Shazam's modus operandi lies in its ability to recognise a single piece of music under diverse conditions. The same track will sound very different when played, with no outside interference, on high-quality equipment, to listening in (say) a car against some engine rumble, to hearing it as background entertainment in a noisy public place such as a shop, bar or cafe. Converting and comparing to a standard format (MP3, for example), then comparing bits, will fail entirely.

Instead, Shazam detects simpler, quality-invariant features of the music such as the tempo, peak rate energy, or number of times the audio signal, across different frequencies, crosses various spectral points, etc. Although these invariants prove to be a more effective approach than bitwise comparison, two points are immediately obvious ... and important:

- It is highly unlikely that any of these features can be detected/measured/recorded perfectly
- No single feature, in isolation, is going to be remotely sufficient to identify the piece uniquely

In other words, a simple, one-dimensional approach will not work; and yet Shazam *does*. Instead, it combines several of these imperfect invariant features, as best it can, into an 'acoustic fingerprint', which—if constructed effectively—may uniquely identify the track. (A fundamental principle of combining datasets in big data analytics is that increased data dimensionality decreases anonymity, whatever the subject.) This acoustic fingerprint can then be sent from the device and queried against an Internet-based lookup (database). Information on the matching music is then returned to the device and offered to the user. The essential components of such a system are therefore:

- As accurately as possible, *and yet imperfectly*, collect invariant identifying features of the music
- Combine these individual features into a single (hopefully unique) acoustic fingerprint
- Transmit this fingerprint and query against a global database
- Return the matched result and all available related information to the user

The mathematics (transforms, etc.) of the construction of the acoustic fingerprint are an unnecessary distraction from our discussion. Suffice to note that it works: the result is sufficiently discriminatory to identify the piece and that this is further made possible by large proprietary databases owned by, or available to, the system as a whole. With this in place, the final step of returning the result and any relevant associated information is trivial.

The reason for this established comparison should be clear, because an obvious question then is how viable such a system could be for ***people***? At a high level, the conversion of Shazam's operation to a form of '*Shazam for People*' (*SfP*) [2] is simple enough in theory, but each step poses questions and challenges in practice. However, here is the obvious initial attempt:

- As accurately as possible, collect identifying features of the person in question. *How? What features might be available?*
- Combine these features into a single '*personal identification mark*' (*PIM*). [To reuse the term 'fingerprint' in this context might confuse.] *Will the result be sufficiently discriminatory? Can it be unique?*
- Transmit the PIM and query against a global database. *Is there/can there be such a database for people? (Or is one needed?)*
- Return identification and all available information to the user. *What parts of this are legal/illegal? Realistically, how effectively could it be prevented?*

We now consider each component of the SfP process in detail.

3. Identifying Features

Considering the principles, established above, that no single feature need be captured perfectly, or can be expected to act as a sole means of identification, several techniques are credible as individual contributions to a PIM. Each of the following recognition techniques is, at worst, an area of active research and many are well-developed in military, security, biometric, commercial, etc. spheres. Each potentially serves as a credible '*identification vector*' (*IV*).

- Face recognition
- Gait analysis
- Body size/shape/proportion detection
- Voice, pitch, tone, language, dialect, accent, etc.
- Chemical/biological/medical analysis (e.g., breath composition, breathing rate, pulse, blood pressure, electro-galvanic skin properties)
- Special characteristics (e.g., scars, injuries, tattoos, piercings)
- Corrective/enhancement technology (currently glasses, lenses, hearing aids, etc. but more advanced 'implants' in time?)
- Unique biometric identification where available (e.g., retina patterns, 'conventional' fingerprints, DNA)

Each of these, inaccurate and insufficient individually, may form a useful component of a compound PIM. However, the concept can be taken further: for additional IVs, there may be situational/contextual data available of comparable value:

- Location (where they are, where they have come from, where they are going)
- Association (who they are with, or talking to)
- Occupation (what they are doing, reading, watching, saying, using, etc.)
- Appearance (what they are wearing, carrying, etc.)

See [3] for a fuller discussion of contextual identification for the '*Prof on a Train*' game [4]. Again, these IVs, limited individually, may prove powerful in combination.

- And finally, but potentially very significantly, *any technology they may be carrying (or wearing or, in future perhaps embedded within them).* If interaction with any of it is possible then a particularly useful IV or set of IVs follows.

Which (combination) of these IVs could be captured in practice, of course, would depend on both the technology being used and its context. Gait analysis, as a technique for example, requires movement. Smart glasses, as a technology by comparison, could perhaps detect most visual signals but would require some extension to perform chemical biological or medical analysis. A more sophisticated contact-based approach could combine more of the latter but might need additional output to convey any returned results to the user. Identification of any indicative technology carried would necessitate IoT-level protocol cooperation. A single device capturing all possible IVs may be unrealistic—at least for the immediate future—so, for any given practical subset, will unique identification be possible?

4. Unique Identification

Currently, Shazam's effective database (including those components acquired from, or in cooperation with, third parties) runs to many millions of tracks. Its method for constructing a unique acoustic fingerprint is sufficiently sophisticated to give identification of 'extremely high but unpublished' reliability from a 10 second sample time. Can the proposed SfP's PIM, from its available IVs, be expected to make a *sufficiently accurate* identification among around seven billion people in the world?

Realistically, probably not—at least not yet. The initial problem is less the theoretical numerical challenge, rather the practical one of technological engagement. The world is unequal. Whilst many in its developed regions have already left their digital impression on (say, in simple terms) the Internet and its data, most elsewhere are effectively technologically anonymous. This, however, is both a help and a hindrance to SfP's chances. Until such time (if ever, of course) as all parts of the planet share the related benefits and perils—uses and abuses—of connective technology, those that are excluded increase the chances of identification for those that remain by reducing potential targets and thus the 'odds' of success.

Ultimately, however, the mathematical viability of any SfP system will depend on the range and quality of IVs that can be collected and the efficacy of their combination into a PIM, which in turn depends on the underlying technology available. Every aspect of this improves almost daily. Considering the current rate of technological emergence, development and advancement, if a completely reliable universal approach is unrealistic today, it would be brave to insist it will remain so a few years in the future. As an example, a company called 'Blippar' already promotes a system, informally described as 'Shazam for Faces', capable of identification of around 400,000 'celebrities and public figures' with 99% accuracy, using face recognition alone [5]. There have also been some disturbingly accurate hoaxes [6].

Once again, for the purposes of this simple discussion paper, we omit the mathematics of transforms, etc., turning individual IVs into a PIM. A more interesting challenge, however, lies in the existence, or otherwise, or even the ultimate necessity of, the global database against which PIMs would be matched.

5. Central Databases and Querying

This may be the most difficult component of SfP—and the most interesting discussion. The concept of a central, Internet-based, queryable database (of people) trivially requires two things:

1. The existence of the database itself, and
2. A search/match standard/protocol: presumably the personal identification mark (PIM) specification,

so it may clarify the argument to consider each of these separately. (It does not look like a particularly difficult exercise to join the two together if they exist.)

5.1. A Central Database of People?

Is a digital database of *everyone* possible? Can it be?

Well, not yet; that is fairly clear. As already mentioned, a large fraction of the world's population have no Internet presence in any form whatsoever. But either that will change or our SfP will have no interest in them anyway. So it is reasonable to start with what we have *got*. Where do we already have *partial* human databases? Could they grow to become what SfP needs?

There certainly are partial DBs already. For 'notables', there is Wikipedia (and worse!); for academics, there is Google Scholar [7]; and many similar platforms for restricted coverage of other areas. And for everyone else, there are social media (Facebook, Twitter, etc.) profiles. Some of these are public, some private, others configurable to be something in between. Some use direct input from the 'person of interest', some do not. Between them, they may largely cover all the ground needed; but, by-and-large, they still have one thing in common: they are all *legal*.

But there is another type emerging . . . Just as an example, consider *Prabook* [8]; then, search for the author of this paper [9]. The page contains a complete potted history of the individual's career and some very personal details too (including parents, spouse, children, key dates, etc.). *None* of this was supplied by the individual: it has all been scraped from other Internet sources, including archive documents in several places. This is not a Wiki of famous people: it is potentially the start of a DB of *anyone*. Prabook claims a mission *'to record and preserve information on individuals who have made a contribution to their nation, local community or any professional field, and on whom sufficient data can be found in books, magazines, public and private libraries, and archives'* but already its motives are being questioned [10].

Is Prabook legal? It almost does not matter because similar sites have appeared and disappeared in recent years: as one is 'taken down' following complaints, another appears. As 'personal information density' increases over the next few years, it is likely—probably inevitable—that, at any given point in time, there will be *something* to target, and these DBs will gradually expand to include more and more people. Like bogus, pay-as-you go 'Who's Who' entries, we can all be famous if someone somewhere profits from it!

5.2. A Personal Identification Mark (PIM) Standard?

On face value, this could be the most difficult component of all. SfP's PIM will require an agreed data standard/protocol: a mechanism for combining the various IVs into a single record, but flexible enough to deal with variation in what features are available in any 'capture instance'. From a technical perspective, this is not hard: separately, and in combination, it is what Shazan and existing face recognition systems (sort of) already do. But surely, unauthorised use of the PIM could be made *illegal*? Surely, any websites carrying a PIM and offering SfP matching services could be taken down by 'the authorities'? Surely, a product or an app offering SfP would not be allowed in the electronic stores?

But it is never as simple as that. Apart from the same problem, as in 5.1, that chasing down these websites is a battle that may never be won, we have the recurrent issue that such technology would have obvious benefits elsewhere (and presumably with individuals' consent). (The same arguments as with sex robot technology [11].) A group of volunteers in an organisation may well want to cooperate in this form of mutual identification and welcome the use of devices, apps and websites using their PIMs for such a purpose. Making a complete data standard/protocol outright *illegal* will be difficult: it has led to contentious debate many times before [12,13].

The upshot (of 5.1 and 5.2) is that, whilst restrictive legislation may be *possible*, it could be hard to enforce. The combination of technology with legitimate alternative use and the practicalities of identifying and dealing with offenders may be too much. (And we have not even mentioned the dark web in any of this!) If there is profit to be made, someone is likely to try to do it. And the fundamental rule of Internet data still applies: if anyone plays free and easy with your personal information, it *may* be possible to trace the culprits, it may be possible to prosecute, even punish, the wrongdoers ... but the damage has already been done: the information is already 'out there'!

6. Returning Personal Information

This remains trivial in a technical sense: once an individual has been identified, any information held in the central database can be returned immediately. But this may be only a fraction of what *could* be available across the wider Internet. The central record would also contain an acquired ('learned') set of effective (combined) search terms that could be used to scrape or mine personal information in real time. Independent data sets could be processed together to de-anonymise material and achieve further identification. If necessary, conflicting records could be 'data-cleansed' and the results iterated back to improve system performance. This might be a critical observation: *once a crude SfP skeleton is in place, it will improve automatically and rapidly*.

Once more of course, a more valid objection relates to legality. Whilst ethical and moral concerns are easily dismissed (they have tended to be historically when there is profit to be made), the law itself is harder to bypass with impunity. But, for legislation to provide an effective check to our proposed SfP, different questions have to be considered:

- What existing (e.g., GDPR) legislation is in place relating to SfP? *Is it sufficient or do parts need to be extended?*
- In whose interests would existing/future legislation be applied? *Who is perceived as requiring protection?*

- Which aspects of SfP are (or could be made) illegal? (Such questions are often complicated by the same technology having beneficial applications elsewhere.) *How effectively, in practice, could any legislation be implemented and upheld?*
- Can privacy legislation ever cope with situations in which no actual data ever exists: rather everything is constructed, combined and processed in real time, then released when done? *Similarly, if different actors were responsible at different stages of the process, which would have broken the law?*

We do not pursue these questions in depth in this discussion paper. Such arguments would be too wide-ranging and/or lengthy for an editorial of this nature. Nor do we consider any (even obvious) knock-on effects of SfP becoming reality (such as making identity theft much easier, etc.). However, the Special Issue that this leader introduces welcomes focused or wider contributions containing legal (or political, economic, ethical, etc.) debate as much as technological content.

7. Conclusions: Putting It All Together (or Pulling It All Apart?)

Futurology is difficult [14]. It is not ultimately clear how an SfP system would work, although there are numerous ways in which, in a few years, it *might*. If a prototype wearable device or mobile app (say) were to be available in five years or so, it could employ *any* of the techniques for feature (IV) extraction discussed here, the exact combination to be determined by the pace and success of hardware and software evolution in each domain. There might even be a particularly disruptive technology on the horizon that could render several of these redundant and make the whole SfP notion even more realistic. Timescales may be debated but the fundamental principle seems sound.

The message, for now, is that such a system is *credible*. There are technical, legal and ethical challenges to consider but none appear utterly insurmountable if there is a will—for good or bad—to make it work. Whether, in practice, such a system does emerge is ultimately not a technological question: those problems are easily solved with time. Instead, whether or not SfP appears—and exactly what it might be used for—are questions whose answers will be determined via the conflicting pressures of profit and public interest. There is little doubt that it would be commercially viable but who would really benefit (or suffer) from it? Will concerns about individual privacy and exploitation, and their political influence, prove to be sufficient and effective restraints? *It could be argued that history warns us against complacency in such matters.*

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Name Any Song in Seconds. Available online: https://www.shazam.com/gb (accessed on 4 January 2019).
2. Grout, V. Shazam for People. *J. Robot Autom.* **2014**, *1*, 1–4.
3. Grout, V. Identity Voyeurism. *BCS IT Now* **2015**, *57*, 24–27. [CrossRef]
4. The 'Prof on a Train' Game. Available online: https://vicgrout.net/2015/08/23/the-prof-on-a-train-game/ (accessed on 4 January 2019).
5. The Shazam for Faces: APP that Identifies People. Available online: https://www.pressreader.com/india/mail-today/20180107/281681140270463 (accessed on 4 January 2019).
6. Facezam Face-Recognising APP Is a Hoax—But It's Disturbingly Close to Reality. Available online: https://metro.co.uk/2017/03/20/facezam-face-recognising-app-is-a-hoax-but-its-disturbingly-close-to-reality-6522373/ (accessed on 4 January 2019).
7. Google Scholar. Available online: https://scholar.google.co.uk/ (accessed on 4 January 2019).
8. Prabook. Available online: https://prabook.com/web/home.html (accessed on 4 January 2019).
9. Vic Grout: Computer Scientist Researcher Educator. Available online: https://prabook.com/web/vic.grout/163108 (accessed on 4 January 2019).
10. Prabook.org—Identity Theft R Us. Available online: http://catanova.blogspot.com/2015/05/prabookorg-identity-theft-r-us.html (accessed on 4 January 2019).

11. Grout, V. Robot Sex: Ethics and Morality. *Lovotics* **2015**, *3*, 1. [CrossRef]
12. Iodine. Available online: https://code.kryo.se/iodine/?fbclid=IwAR0PlTzBeohLylafTvY-zrIP7Wb5xE4BCuHxIV4EjFVfXpJTw56meubddxo (accessed on 4 January 2019).
13. BubbleStorm: Rendezvous Theory in Unstructured Peer-to-Peer Search. Available online: http://tubiblio.ulb.tu-darmstadt.de/74907/?fbclid=IwAR0A4XHd8mNQpekjr8dST64ICwKr2D_KngjN-kt5JuvJO5PQcKLrVSziYrI (accessed on 4 January 2019).
14. The Problem with 'Futurology'. Available online: https://vicgrout.net/2013/09/20/the-problem-with-futurology/ (accessed on 4 January 2019).