AN INTERNET PROTOCOL ADDRESS IS NOT A PERSONAL IDENTIFIER

As it has become more popular, many misconceptions about the Internet have crept into public consciousness. Because it is large, complex, and incomprehensible from any single point-of-view, many people make the mistake of thinking that there is some kind of technical mastery behind it all, to make it all work. Nothing could be further from the truth. The Internet is messy, error-prone, and inefficient. It was designed right from its origins to be thus — "assume the network is unreliable" was a mantra of its early developers. One part of the Internet doesn't know what another part is doing or what it is like, and it doesn't need to know this information to function. Limited and imperfect knowledge is all that is required for communication to succeed. The Internet is a machine language for communication, and just like a natural human language it has rules of grammar and spelling and pronunciation that are broken all the time by its native speakers.

An Internet Protocol address, or IP address, is a sequence of ones and zeroes that identifies a node on a network. A node is a computer or a smartphone or some kind of intelligent device that can communicate. Knowing this address, an intelligent communicating device can calculate how to move data farther along in the network towards the final destination, and ultimately right up to the destination itself. That is a procedure called routing. IP addresses are the information, and routing is the action based upon this information.

Notice that in talking about an IP address I didn't say anything about a human being who might be using the node on the network. I didn't need to, because an IP address doesn't have anything to do with a person. An IP address pertains to the task of networking devices.

The Internet works according to a layered model. Layers or functions define a type of action that must be taken for communication to occur. There are identifiers that are meaningful in describing what is going on at each layer. An IP address is an identifier for a node, and it is meaningful at the Internet layer that defines how packets are routed to their final destination. But that is just one of the functions that must be carried out for successful communication to occur.

There is no authentication between layers. This is referred to as a “stateless” mode of communication. Think of it like the connection between a person dropping a letter in a mailbox and the post office worker who collects all the letters to take to the sorting station. What is the connection between these two steps in communication? There isn’t one, immediately. Each function can operate while treating logically prior functions as a fait accompli. I can drop off my letter with only the vague awareness that it should be picked up later; the post office worker can collect the letter with only the vague awareness that it is properly addressed and has the right postage.

With the Internet, there is no direct connection between a packet of data moving across a network and a human being who may have initiated the sending of that data. That is because the Internet developed from earlier “store-and-forward” models of communication, that worked much like the post office in delivering paper mail.

Network communication can be initiated either by a human being operating a computer, or by the computer itself. From the point of view of the network, there is no way to know the difference. That is like saying that the post office can deliver mail where it was individually sent by a human being or sent in bulk by a machine, or it is like saying that the telephone company can route a call whether a human being is calling or if a machine is robo-dialing the call.

In summary, the layered model of the Internet can be described as follows. A process runs on a computer, and may be something like browsing a web page or sending an email. The “conversation” between computers at this layer can be described as a session, and it has identifying descriptors including port numbers. A packet of data containing a very small part of the conversation is then transferred to the networking layer, where an identifying number called an IP address is attached to it, to help route this piece of the conversation to its final destination. Now properly addressed, this chunk of data is passed to the data link layer, where it is goes out an interface or port to exit the computer to traverse a network, where it will transit a series of interfaces on its way to the final destination. There is an identifier at this level too, and it is commonly an Ethernet address that is associated with an interface. Finally, a signal is sent across a wire or through the air to physically reach another interface. Once at the final destination, the entire procedure is followed in reverse, until the corresponding application at the other end receives this small packet of data, which is one piece among many of the overall conversation.

An IP address is not a personal identifier, and on its own cannot connect a human being to data transmitted across a network. The Internet does not have such a thing as a personal ID, like a Canadian has a Social Insurance Number in interactions with the federal government. The only way to connect a person with Internet use is by taking a bottom-up approach to the network layers, and capturing all of the meta-data and data involved along the way.

Here is what it would take to convincingly prove that a person did anything on the Internet. First, prove that a person made use of a networked device by possessing and operating it. Then ascertain the identifying information at the data link layer for the port that sent and received packets of data. Then connect this meta-data with the IP address used to identify the node in the whole Internet. Then isolate the session layer information that identifies the particular conversation. Finally, reconstitute the pieces of the conversation, such that the actual data transmitted or received can be determined.

When the Internet began 40 years ago, each of the layers had a physical analogue and it was easy to picture what was going on. An operator was a real person; a node was a real computer; a port was a physical connection with a wire attached and there was only one address needed to identify it uniquely. Now, the trend of virtualization means that every level is more abstract. A node can be a virtual machine, which can be one of dozens or hundreds of computing/communicating computer-like objects inside the physical box. The physical port can manifest itself as any number of network presences, each one of which can have many IP addresses associated with it. From a static and human-shaped Internet, we have evolved to a dynamic Internet consisting mostly of machines talking to machines.

The amount of information that must be collected to tie an Internet Protocol address to the activities of a person using the Internet is very large, and it is computationally and monetarily expensive to record and store all of this. Logging data and meta-data, on its own, serves no security purpose. Only analyzing logged data serves a security purpose. This too has costs, which are substantial.

It is technically possible to record every telephone conversation. Capturing, storing, retrieving and analyzing all of this information would have large costs. It is technically possible to open everybody’s mail. Recording all the information contained in postal addresses and letters and then analyzing all this meta-data and data would have large costs. For the Internet, large data centres would have to be maintained to store all the captured meta-data and data, and large numbers of skilled analysts would have to be employed to make sense of it all.

The replacement of IPv4 with IPv6 does not change the nature of the Internet or the purpose of the Internet layer. The Internet is still a “best possible effort” relay system, and the Internet layer is still stateless, which is to say it does not have error corrections at this stage.

It is not true that IPv6 addresses are “persistent” in a way that IPv4 addresses are not. In fact, the opposite is the case. IPv6 addresses are transient, and more removed from human agency. When IPv4 addresses were introduced in 1982, they were fixed, unique addresses assigned by human beings to networked devices. As IPv4 evolved, ways of assigning addresses that were more automated began to be used, such as the Dynamic Host Configuration Protocol. With DHCP, pools of addresses are controlled by an Internet Service Provider, and assigned to connecting devices and revoked from disconnecting devices without human intervention. Now, with IPv6, we have a built-in scheme of stateless auto-configuration, where the device itself “makes up” its own interface identity, solicits a network prefix from a router, and concocts a complete 128-bit address thereby. An IPv6 address is a unique identity among 380 undecillion possibilities, and is firmly in the realm of autonomous technology — it is all about machines talking to machines. The future shape of the Internet is such that IPv6 addresses not only can be but perhaps should be random, unpredictable, and utterly removed from human agency.

If one has in mind a single IP address fixed to one physical device, which is proven to be under the control of one identifiable human being, then one is thinking about the Internet as it was created in 1982. The Internet does not work that way now, and it will be even more removed from that in the future as it improves in scale with IPv6 and improves in security with strong encryption and authentication.

One might consider consolidating multiple sources of information. The Internet is a node-based network of networks, so let’s suppose we gather meta-data and data anywhere along the line from the source, through intermediate relays, to the final destination. The “store-and-forward” nature of the connection between the nodes is stateless, just like the connection between the layers in the network stack is stateless. Therefore, it is up to us to put the pieces together. A unique problem with coordinating information gathered from different nodes arises from the need to accurately establish a timeline. The “chain of custody” must be followed through time. Timekeeping with most computing and networking devices is not very good. The most discrete interval most devices comprehend is a millisecond. A gigabit network card can push packets across a wire at a rate of over 14,000 per second. That is a rate 14 times greater than the most granular measurement of time built in to most computers. The Internet has a feature called the Network Time Protocol, but there is no requirement to run it. Most networked devices are going to have the wrong time, uncoordinated time, and have a coarse measurement of time. Think of it like a grainy surveillance video in a store — it may record a robbery, but it may be unusable for the purposes of an investigation or a prosecution.

This is not to say that gathering information from many sources is impossible. It is problematic, though, and it is not a panacea to the substantial cost and effort required in order to convincingly tie an Internet Protocol address to the actions of a human being.

Trying to isolate Internet spying to a domestic context is impossible, because the Internet is by its very nature trans-national. The trends of virtualization and so-called “cloud computing” make this even more so. The connection between an IP address and a human being is a weak one, and the always indirect connection between what happens at the network level and what a person does has been made even more tenuous and remote by the explosive growth and speed of the 21st century Internet.

Source : https://www.exitcertified.com/blog/michael/2013/an-internet-protocol-address-is-not-a-personal-identifier/