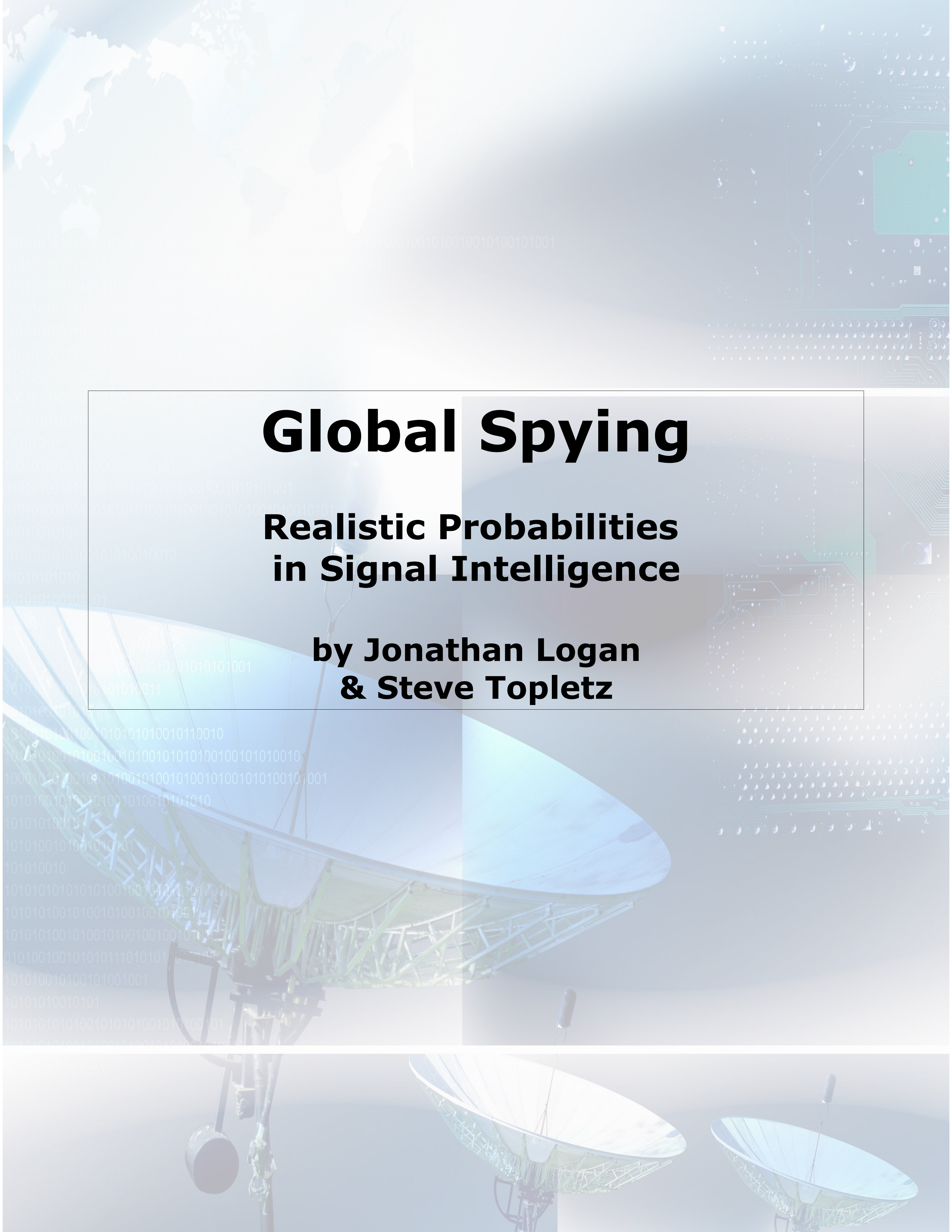

Global Spying

Realistic Probabilities in Signal Intelligence

by Jonathan Logan
& Steve Topletz

PREFACE

In this article I will present insight into the realistic possibilities of Internet mass surveillance. When talking about the threat of Internet surveillance the argument most often presented is that "there is so much traffic that any one conversation or email won't be picked up unless there is reason to suspect those concerned; it is impossible that 'they' listen to us all".

This argument assumes that there is a scarcity of resources and motivation required for mass surveillance. The truth is that motivation and resources are directly connected. If the resources are inexpensive enough then the motivations present are sufficient to use them. This is visible in the economic effect of supply availability increasing demand. The effect is that since it is more easily done, it will be done more readily. Another fault in this argument assumes that there is only all-or-nothing surveillance, which is incorrect.

INDEX

I. RESOURCE REQUIREMENTS

It is important to break down the resources required and methods available as well as the means of surveillance to understand what realistic threat mass surveillance of digital communication is. The resources required are Access, Storage, Bandwidth, and Analysis. In this paper, I am speaking about digital communications, and these methods do not fully apply to purely analog communication, such as POTS (normal telephone service).

ACCESS

Surveillence requires acces to the communication to be surveilled. Data today is transmitted via copper cable lines, fiber-optics, directed micro-wave communication, broadcast radio methods (WIMAX etc.), satellite, and a few other arcane methods . The most profitable transmission media for surveillance, by far, are fiber, broadcast, directed micro-wave, and satellite. Fiber provides the benefit of large amount of data from a single “cable.” Broadcast radio provides the benefit of non-physical accessibility. Directed micro-wave is easily acquired through classic stand-in-the-middle listening. Satellite provides a very big footprint, where one needs only to be standing near the receiver of the transmission.

Fiber cables provide the most interesting targets for surveillance. Almost all international communication eventually goes over a few particular fiber lines, so this is where the tapping is focused. This is a practice far different from the UK / USA Echelon system of the 1990s which operated mostly by targeting direct micro-wave and satellite transmissions, because international fiber-optic lines were more rare. Today, tapping into fiber is easily accomplished through a variety of methods: splicing the fiber-optic line, or connecting to the repeaters, or tapping into the endpoint routers and even more esoteric methods like bending the fiber and detecting stray "ghost" photons[1]. Tapping in most cases is purely passive, which means two things. First, the signals are being listened to and not intercepted or modified. Second, signals are non-trivial to detect by the endpoint, which means there is no *click* on the phone to tell you that someone is listening in.

Access to fiber-optic lines is mostly accomplished by connecting to repeaters and tapping endpoint routers. That is what is being performed by AT&T at the request of the NSA. This method is inexpensive in resources, easy to implement, and requires very few people to know and operate. In the case of repeater connections, even the fiber owner may not be aware unless they find the tap during routine maintenance.

Civilians generally assume that the Internet consists of millions of independent lines that would have to be tapped individually for mass surveillance. Luckily for signal intelligence gathering and analysis, this is not the case. To tap into 90% of traffic connecting the Eastern Hemisphere to the Western Hemisphere (GUS / RUS / AFRICA / MIDDLE EAST / EU to US), agencies only need access to either 30 fiber cables[3], or half of the 45 landing points[4]. An alternate method to achieve such access to this traffic is to install access devices in just 7 of the correct IXs[5] (Internet Exchanges), which is where ISPs and backbones interconnect at a single location. Rest assured, all of above has happened at various scales.

A special property of the Internet, which lends itself to accessibility, is resiliency in routing: if you can not tap into a specific route then you can destroy it to have the traffic rerouted through lines that you have full access to. Accidently drop an anchor on a submarine cable, or have an excavator accidentally cut a line, and then execute a Distributed Denial of Service or Table Poison attack against the routers in question. There are an endless amount of innocuous events which are created or exploited for covert access to fiber-optic communications. For example, an event occurred in October of 2007 where the CAT5 cable between Iceland and Scotland was apparently severed, rerouting all traffic through the USA. Such an event could easily be used for purposeful traffic rerouting, a tapping opportunity, or both. For tapping subjects that require more surgical precision and shorter time windows than typical dragnet operations, there are additional options like breaking into routers to establish "shadow" routes on IXs and landing points.

It is important to keep in mind that a surveillance organization does not have to cover all nodes or routes for full access. Simply select the ones with the most connections or throughput to other nodes and you have succeeded. Tapping into the connection at *any* endpoint, transmission line, repeater or router is enough to obtain the access required for mass surveillance. After you have access, the remaining work for mass surveillance is relatively trivial.

STORAGE

Storage, as well as bandwidth, are relatively expensive resources. It makes little sense to tap into communication lines and not be able to store the data that you want. However, if you are able to select, reduce, and compress the data you are interested in, then storage resource requirements decrease. The cost of storage today using standard products on the market is high when compared to the total amount of traffic traveling the internet. The cost of storing a years worth of traffic is very high; for 2008 alone it would cost over $51 billion (90 exabytes). However, if you use data reduction methods then the total storage costs are much lower. For example, it is not necessary to store a copy of all traffic each time someone downloads a movie, it is enough to reference the movie. The same applies for webpages, documents, and other uniform communications. By storing only unique internet traffic at the data-mining facility, storage costs are reduced to much less than 1% of the original projection, which brings mass surveillance into close reach for many organizations like the NSA, which has a projected yearly budget[2] between $3.5b and $4b, excluding “blackops”. Italy implemented such a system in 2007, named DRAGON, to retain data acquired from the mass surveillance of their citizens. Some countries, like Sweden, have even decided to record all international traffic crossing their borders.

BANDWIDTH

Captured data must be transferred from the temporary storage on the tapped line, to the aggregate data stored at the data-mining facility. Therefore, data of interest must be transferred to a collection point. Using the above projections, transferring unique traffic from the tapping point to the data mining facility costs roughly $150 million. This is entirely in the financial reach of both large and small intelligence gathering organizations. Although it is not publicly known if any organization does indeed copy and store all unique traffic on the Internet, game theory suggests that if it is both possible and beneficial that not only is it likely, but that capable parties will scramble to do so just to remain on par with their counterparts.

ANALYSIS

Analyzing the stored data is where real intelligence happens, and is more demanding than both storage and bandwidth requirements. Post-tapping analysis and offsite analysis should be differentiated: Post-tapping is what selects and reduces the data to what is unique, offsite analysis is where raw data is turned into intelligence to be acted upon. Post-Tap analysis typically occurs directly at the tap, and the resulting data is stored. Very little communication is of interest for realtime surveillance, so data is rarely relayed immediately, and is typically cached to be transmitted at a time better suited to both the cost and detectability of the surveillance. The cost of post-tap analysis is approximately $4.5k per Gbps of bandwidth. This means that post-tap analysis cost for all unique Internet traffic is roughly $750m per year. Offsite analysis costs varies, and depending on what operations and techniques are performed on the unique data collected from the entire Internet, costs could start at a few million dollars, and reach up to a $1.5b in yearly costs.

The annual cost of Storage, Bandwidth and Analysis requirements are therefore around $2.25b. The cost for the Access requirement is a little lower, likely hovering around $1.5b. Once all of this data has been collected an analyzed by computers, it then requires human resource components to act upon the data, costing an additional estimate of $1.5b. Therefore we arrive at the total cost of surveilling **all** unique internet traffic being around $5.25b, with a variance of around $500m depending on what is done with the information. Since the regions of interest are different with some parties focusing on national rather than global surveillance, the required “black” budget for total mass surveillance of the Internet is less than $1.5b per interested party.

Economically speaking, this is far less then many countries spend on things like military weapons or state police. This financial estimate assumes that the selection of communication is 100%, without regard to protocol, and includes all website, email, and voip traffic. This estimate also assumes that it is a single party doing the work, and that resources like taps, storage, and manpower are not being shared. As we can see in most "developed" countries today the actual work is outsourced by legislative mandate, such as the EU Data Retention Directive, which provides no funding and shifts the burden entirely upon private Data Centers and ISPs. In some countries, ISPs are required to provide the Access, Storage, and Bandwidth components, or do it for their own profit by participating with interested 3rd parties like Phorm or Google. Given the minimal costs compared to both the budgets and perceived benefits, it is naive to assume that mass surveillance is not being employed.

II. METHODS OF POST-TAP AND OFFSITE ANALYSIS

A netflow is a relationship between one computer and another one, the word "connection" does not really apply to packet-based networks. A thousand active “professional” internet users create between 300k-500k concurrent netflows with roughly 80 Mbps to 250 Mbps of sustained bandwidth consumption. Occasional internet users, the majority, create much less. The numbers appear huge at first glance, but applying professional processing equipment and software can reduce those huge numbers to an easy to handle set of information that can readily be acted upon. Communication surveillance analysis uses the Escalation of Surveillance concept, executed by four basic methods: Classification, Interpretation, Reaction and Selection.

Escalation of Surveillance means that depending on previous analysis the computers reserve more resources to spy on a specific target. How they do it depends on the rules given to the Reaction component and can be exceptionally complex. The escalation process does not stop at the post-tap analysis stage but "trickles up" to the offsite analysis. Additionally, if a target becomes interesting due to escalation then other people in connection with the target become more interesting as well. This is because of context classification, and can be summed up as “guilt by association”. Technology makes it possible to seamlessly and inexpensively interconnect the post-tap installations and the semi-automatic creation and updating of reaction rules. Therefore escalatation of resources spent on ancillary target groups that are connected to an escalated target can happen almost in realtime.

When communication is tapped into, the first step for analysis is **Classification**. The two types of classification are *Content* classification and *Context* classification.

Context classification defines what data is transferred, and who transfers it. Context classification on IP networks, such as the Internet, is trivial because the underlying protocols provide all required information in a form that is easy for computers to read and understand. With the advent of Deep Packet Inspection, a popular buzz word around voyeurs, the context classification even touches the application protocols (Layer 7 analysis) and payload (classical deep packet inspection). The result is not just having the conclusion "XY reads a Google page" but being able to state "XY searched for Porn on Google.” The data generated by context classification is ideal for storage and later data-mining. Such data sets are relatively small and have a precise meaning.

Content classification defines what type of data is transferred and what meaning the data has. In most cases content classification only considers the *type* of data, such as pictures or movies, but in some cases the *meaning* of the data is of interest. Content classification is especially effective on unique Internet traffic. The Google logo is transferred a billion times a day, however it is not unique. It is classified once, put into a reference table, and never revisited. The same goes for most web and p2p content. Combined with context classification a resulting data set would say "XY downloaded a nude picture of Angelina Jolie from webpage Z". The resulting dataset will be less then 200 bytes, regardless of picture size, and by the time the first 5 to 10 packets are transferred, the connection has already been analyzed. If the content however is not unique then the Classification method fails and the next method used is Interpretation.

Interpretation of unique data means that the data is translated into a form that data mining can act upon. For email that means that the text/documents are analyzed by language/semantics analyzers. It is a technology that was the big buzz-word 8 years ago and available freely to the market. Such an analyzer running over this article will spit out: "Analysis of internet surveillance feasibility". These tools are able to find out the most important words, places, times, subjects, people mentioned in the communication content. And it takes often only fractions of a second to do that. I am not talking about "hit words" here. That is totally old school. I am talking about finding out the "meaning" of a text. This goes so far that it is used to find out if the text actually includes a code ("the parcel has been delivered to our friend"). The resulting data set for that will be relatively small (around 2-5KB), machine readable and easy to store. After content/context analysis and interpretation are done the result is a data set that can be reacted to.

The next two steps are a dual-factor component, requiring both human presets, and computer processing. **Reaction** is a programmed into the computer as a rule set, as it requires cognitive abilities beyond those that a computer can intuitively choose or measure. In the **Selection** process the data is combined with a vector that holds "points" for the various interests the spying party associates with it. While the programming is a thing done by humans the "interest vector" is attached automatically. Depending on the "interest vector" the data might be thrown away, cached locally to be combined with additional data or transferred to offsite storage & processing. Both Reaction and Selection are completed very quickly, during which the parties of communication are re-classified as well, which accomplishes Escalation.

Computers can make a lot of sense out of seemingly harmless data. They are able to correlate many communication processes, and they are able to remember things of raised interest. Given the low cost of processing required at different stages and the cheap storage available, it is likely that a historically detailed profile of all communication of an individual is created.

III. IMPLICATIONS

The result of inexpensive internet surveillance measures that do not require human intervention is a collection of data for offsite analysis and reaction. It is entirely possible to automatically create classification, interpretation and reaction rules that preselect certain communication participants for more in-depth surveillance without any human interaction.

If a person shows an unusual communication pattern, perhaps at 80th percentile, then this person becomes someone of greater interest to agencies conducting espionage. The communication patterns that are analyzed could be over months, and include online hours, contacts of 1st and 2nd and 3rd degrees, web search terms, and the interpreted content of all your communication. The only thing that effectively keeps spy organizations from automatically spying on you is if your total communication profile, and the communication profile of the people in your social environment, are entirely uninteresting to them both now *and* in the future.

It is feasible and realistic to expect that Internet mass surveillance of a certain scale and reach already exists world wide. The analytic capabilities of current technology is exceptional; and since the long-term memory is inexpensive for data of interest, it is therefore likely to exist. That means that both innocent actions and the actions of those in your social environment can trigger more in-depth surveillance in an automatic fashion. The human and technical resources required for Internet mass surveillance are not only within the reach of many parties, but constitute a small fraction of their available resources. If it is assumed that there is *any* motivation for mass surveillance, then all other factors aside, the economics suggest that it is performed on an astronomical scale not only by nation states and their agencies but also by corporations. Looking at the sales data available for specialized surveillance and analysis equipment offered to the market it is naive to assume that many bytes of communication escape surveillance.

The distinguishing matter is not if individuals are being spied on by computers, because they certainly are, but if they are also being spied on by people. Signal intelligence always has been a large portion of an intelligence agency's budget, and is more so after the American tragedy of September 11th. International corporations that try to control information leakage, public image damage, competitive analysis and outright espionage are also increasing their signal intelligence budget. Furthermore intelligence gathering is the bread-and-butter of many “dot com” companies that provide their services for free, such as Google, Yahoo, and MSN. These companies and their offerings are ubiquitous, so the issue is not *if* or *why* they do it, but *how* you become a person of interest.

IV. THREAT ASSESSMENT

The specific motivation to select your communication for analysis does not have to be high at all. It is an anticipated future interest, and is visible in data retention and other “preventive” measures employed by governments today. The motivation can be anything interesting to an agency, from web searches about tax savings, emails from those with unpopular political opinions, interest into certain technological trends, the layout of your stock portfolio, the grade you achieved in your chemistry course, the position you hold in a company, and participation in a group of interest. The list of “interesting” activities is innumerable, and the more interesting your activities, the more elevated you are as a surveillance target. In fact, anyone reading this paper would almost certainly elevate their status as a surveillance target. Staying below the radar can be extremely hard if you are in any way different then the majority of the populus.

When surveillance becomes trivial for an unrestrained party then it will be done, and sadly there is no good reason that they should not do it if unrestrained. Most of the notions against the reality of mass surveillance are based on the "scarcity of resources and motivation" argument. It has been demonstrated that there is no scarcity of resources to do surveillance or store its results, only to act upon it by human resources. In our current world, there is no scarcity of motivation to do it. In fact, there is a whole industry and even political parties lobbying on the behalf of surveillance. There are enough power-hungry people that want to stay in power, and institutions that create a life of their own. Someone once said that the Internet is not only the best tool for mass communication but also the best tool for mass surveillance and control ever created. That person is right.

VI. END NOTES

This article exclusively deals with the possibilities and methods for passive surveillance of non-participants of the communication being surveilled. There are numerous other methods of surveillance and data collection existing on the Internet. Those include cookies, spyware, log file aggregation, system fingerprinting, and many other methods.

VII. Q&A

Q: What about using word scrambling to defeat language analysis?

A: The technology used in most word processors is good enough to instantly reconstruct large portions of a scrambled text. Systems working with semantic analysis, context and subject discovery as well as whole text probability approaches are even better. They might not be able to reconstruct every single word, but enough of the content to make sense of it. The same is true for most if not all "good advice" given by friends. Good security is not that easy. If advice does not include strong cryptography, it is uninformed at best, and disinformation at worst.

Q: Are encryption users more likely to become targets?

A: As mentioned in the article one of the methods used is to find out unusual traffic and content patterns. Using email encryption is something unusual for the normal population. There have been several cases where the use of encryption increased the interest of investigating agencies. However I still think that it is a necessary and smart thing to encrypt everything you can. Surely you cannot beat context analysis with encryption alone, but content analysis and interpretation can be made much less effective or even impossible.

The advise I would give is to encrypt all your communication every time. It is better to have a consistent communication pattern than to only encrypt occasionally because the total amount of valuable data collected will be lower. If you are only encrypting information you think is sensitive, then it is also known which communications should be more heavily analyzed.

Q: Are people using anonymity networks more likely to become targets?

A: Yes. The total number of available anonymization services is small. Just a few thousand computers in total are serving in publicly available anonymity networks. To target all traffic going to or from those computers is trivial. However only a really big adversary will be able to automatically trace and connect the various relayed packets to each other, and those adversaries surely exist.

Looking at the network layouts of the more popular anonymization networks it is actually not hard to watch all traffic they relay. Some services make it hard to identify single communication events when watching only a limited set of the total connections that exist at the same time by increasing the crowding effect (hiding in the crowd). With effectively executed crowding, you will be seen but not necessarily identified.

Q: But company X said they use technology Y, won't that protect me from all adversaries?

A: No. It is true that technologies exist to drastically increase your privacy on the Internet. However, none of them protect you against an omnipotent attacker. Most are good for evading nosy marketing groups, few are good enough to hide yourself from the eyes of domestic security agencies. However, none will protect you against a motivated attacker with global access to the Internet. If your anonymization service is decent then they will have a note in their website or documentation that effectively states "do not rely on this technology if you require strong anonymity." If they aren't they will say: "we make you 100% anonymous on the Internet."

Q: What can be done?

A: Writing to your congressional representive will not stop spying. Politics and public opinion will not help at all to reduce or even solve this problem, because politics and public naivety created the problem. There are only four things you can effectively do:

1. Accept that the world is *not* a place where everyone believes others should be free.
2. Use technological self defense such as adequate anonymity services and best practices.
3. Get others to defend themselves as well.
4. Fight against any force that wants you to give up your freedoms and privacy.

Protecting your privacy does not come for free today, and it never has. One last word to the wise:

Those that shout the loudest that they will protect you or those that do it for free are not necessarily those that only have your freedom and privacy in mind: There Ain't No Such Thing As A Free Lunch

VIII. ABOUT THE AUTHORS

Jonathan Logan works as a communication network consultant for Xero Networks AG and Cryptohippie PA Inc. He can be reached via email at j.logan at cryptohippie.net (PGP Key: 0xE82210E6) Steve Topletz is the operations advisor for XeroBank, a group operated by Xero Networks AG. The opinions expressed in this article are those of the author and do not reflect the views of Xero Networks AG, or Cryptohippie PA Inc., their management, or their respective owners. If you want to distribute this article please contact the author.

1. http://blogs.techrepublic.com.com/security/?p=222&tag=nl.e036

2. http://www.fas.org/irp/commission/budget.htm

3. http://www.telegeography.com/ee/free_resources/figures/ib-04.php

4. http://www.telegeography.com/ee/free_resources/figures/ib-02.php

5. http://en.wikipedia.org/wiki/List_of_Internet_exchange_points_by_size