STEWART BAKER'S IM-Y NUMBERS

Stewart Baker accuses Bart Gellman and colleagues of inventing a phony statistic when they note that 89% of the communications collected under Section 702 were non-targets. He does some math to prove why they're wrong in their interpretation of the scope of this.

> The story is built around the implied claim that 90% of NSA intercept data is about innocent people. I think the statistic is a phony. Especially in an article that later holds up US law enforcement practice as a superior model.
>
> What's wrong with the statistic? Well, let's take an example from law enforcement. Suppose I become the target of a government investigation. The government gets a warrant and seizes a year's worth of my email. Looking at my email patterns, that's about 35,000 messages. About twenty percent – say 7500 –are one-off messages that I can handle with a short reply (or by ignoring the message). Either way, I'll never hear from that person again. And maybe a quarter are from about 500 people I hear from at least once a week. The remainder are a mix – people I trade emails with for a while and then

> stop, or infrequent correspondents that can show up any time. Conservatively, let's say that about 25 people are responsible for the portion of my annual correspondence that falls into that category. In sum, the total number of correspondents in my stored email is 7500+500+25 = 8000 or so. So the criminal investigators who seized and stored my messages from me, their investigative target, and over 8000 people who aren't targets.
>
> Or, as the Washington Post might put it "7999 out of 8000 account holders found in a large cache of communications seized by law enforcement were not the intended surveillance target but were caught in a net the investigators had cast for somebody else."

I agree that the numbers would be impressive — if they actually were what Baker claims they are.

But they aren't.

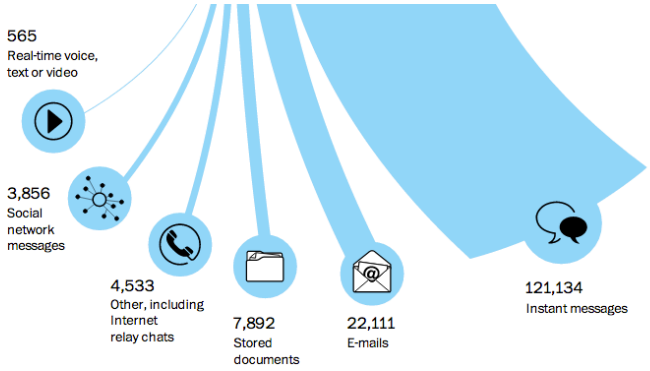
First, remember that these are minimized communications. And while the NSA is keeping data that has no foreign intelligence value, it is almost certainly not keeping spam (we know this because other NSA documents talk about defeating spam). So eliminate that 20% — or likely higher — or so right off.

Furthermore, the 9/10 ratio does not reflect all the communications WaPo examined. It doesn't include the minimized US person ones.
Almost half of the communications NSA identified as US person communications — that's somewhat clear from the graphics, but Gellman stated that on Twitter.

So the actual number is closer to 95% of communications not being targets, not 89%.

But Baker also doesn't consider what he's dealing with. For the most part it's not email,

it’s IMs.



76% of this sample is IMs. Just 14% are emails.

So while Baker’s email example is nifty, it’s largely off point. Because he’d need to look at his IM patterns (or those of a 25 year old, who is more likely to resemble a target), not his email patterns.

It would still be a low number, if you’re considering pre-processed communications. It makes more sense when you realize that’s not what you’re considering.