# Inteligência Cibernética e a Linguística Forense como ferramenta - o uso da análise linguística para atribuição de autoria em ciberataques

Por Leonardo Perin Vichi *

Em 12 de maio de 2017, o mundo assistiu perplexo à propagação de um violento *crypto-ransomware* que varreu o ciberespaço, sequestrando computadores ao redor de todo o planeta. O *malware* ficou conhecido com o nome de *WannaCry*. Este ataque por *ransomware* - modalidade de ciberataque que criptografa o conteúdo de computadores, bloqueando o acesso a este conteúdo e exigindo resgate em dinheiro -, paralisou empresas, órgãos governamentais, bancos, operadoras de telefonia, como a *Telefónica*, na Espanha, e atingiu até mesmo hospitais que necessitaram paralisar cirurgias ao se verem sem acesso ao prontuário de seus pacientes.

Nunca a Rede Mundial de computadores esteve tão perto de algo semelhante a uma Guerra Cibernética quanto durante este ataque que revelou a fragilidade de nossas infraestruturas críticas de informação e a facilidade que um simples código mal intencionado tem de causar efeitos devastadores em escala planetária.

Ataques do tipo se tornaram extremamente frequentes e a cada dia novas modalidades parecidas são registradas. No Brasil, os exemplos mais relevantes de ataques por *Ransomware* foram o sequestro dos computadores da ANATEL pelo grupo cibermilitante *Anonymous* em 2016 e o sequestro dos computadores da prefeitura de Pratânia, no interior do Estado de São Paulo, em 2015. A atribuição de autoria a esses ataques é um desafio para empresas de Segurança da Informação e órgãos governamentais de Defesa Cibernética. Considerando que seu código-fonte muitas vezes é reaproveitado de outros códigos-fontes similares e que as contas de pagamento por *blockchain* usadas para cobrança dos resgates não permitem identificação de seus proprietários, poucos são os dados que facultam o trabalho analítico em busca da identificação dos perpetradores de tais ataques cibernéticos. Contudo, ataques do tipo apresentam alto risco de provocar disrupção de amplo espectro, atingindo infraestruturas críticas, paralisando serviços essenciais ao funcionamento do país, e, por isso, exigem o uso de ferramentas cada vez mais sofisticadas, não apenas para detecção de tais ameaças, mas para a investigação em busca de seus autores.

Nesse sentido, a atribuição de autoria em ataques por *Ransomware* possui dois objetivos primordiais. O primeiro, em especial, em ataques que exigem resgate em dinheiro, é o de dar segmento a investigações policiais para o devido encaminhamento do processo legal, visando a responsabilização criminal dos envolvidos, ou seja, a condução de investigações dentro do escopo das delegacias de crimes cibernéticos. Um segundo objetivo apresenta questões estratégicas sub-reptícias de maior monta. Independentemente de haver exigência de pagamento de resgastes, estes ataques podem trazer motivações ligadas aos interesses de atores estatais que visem testar a resiliência ou até mesmo desestabilizar a infraestrutura informacional de um determinado país. Portanto, é questão que se relaciona com as competências e interesses tanto da Segurança da Informação quanto da Defesa Cibernética.

Se dados como códigos-fontes, IPs e contas *blockchain* não permitem muito avanço na atribuição de autoria em ataques por *ransomware*, o manifesto de pedido de resgate traz informações que permitem análise linguística e perfilamento de seus autores.

O uso da análise linguística e da Linguística Forense para perfilamento de criminosos não é recente, trazendo em sua história casos famosos com o de Ted Kaczynski, o *Unabomber*, Derek Bentley e Timothy Evans. A construção de perfis linguísticos é realizada através do emprego de métodos da Linguística Aplicada através dos quais conjuntos de marcadores individuais e únicos de uma pessoa podem ser detectados, seja para depreensão de informações acerca de sua idade, escolaridade, classe social, gênero etc., seja para comparação entre evidências para esclarecimento de autorias contestadas.

**A Linguística Forense para identificação de autoria foi utilizada com sucesso em diversos casos de ciberataques, inclusive no ataque do *Ransomware WannaCry*.**

A Linguística Forense para identificação de autoria foi utilizada com sucesso em diversos casos de ciberataques, inclusive no ataque do *Ransomware WannaCry*. Em 2018, os laboratórios da Kaspersky detectaram que o código-fonte do *malware* financeiro Karamanak/Pegasus/Ratopak foi escrito por falantes nativos de russo, o que corroborava a desconfiança de que o código malicioso tinha como alvo instituições financeiras daquele país.

No caso do *ransomware Wannacry*, este trazia o pedido de resgate redigido em 28 idiomas diferentes, entre eles o Chinês (tanto o simplificado quanto o tradicional), Finlandês, Holandês, Inglês, Francês, Alemão, Indonésio, Italiano, Japonês, Coreano, Norueguês, Português, Romeno, Russo, Espanhol, Sueco e Turco. Pesquisadores da Flashpoint, empresa de Inteligência Cibernética com sede em Nova Iorque, nos Estados Unidos da América, analisaram os 28 textos e constataram que o nível de correção textual do texto em Chinês, seja pela pontuação, sintaxe e nível gramatical empregados, indicava que o texto fora redigido por falantes nativos do idioma e escritos em um teclado próprio para o idioma chinês. A versão em inglês do texto trazia agramaticalidades que indicavam que o falante não era nativo do idioma ou era possivelmente alguém que recebera educação formal de baixa qualidade. Vale destacar aqui que o nível e a qualidade das agramaticalidades produzidas por falantes nativos e falantes do idioma como segunda língua são diferentes e dificilmente se confundem. Os textos em outros idiomas aparentemente, segundo a empresa, teriam sido produzidos por tradução assistida por computador, apresentando cerca de 96% de correlação com as traduções realizadas pelo laboratório usando o Google Tradutor.

Entretanto, a análise pericial em Linguística Forense pode ser prejudicada por contramedidas que visem lançar uma cortina de fumaça sobre as evidências para atrapalhar quaisquer tentativas de atribuição de autoria linguística. Métodos como passar o próprio texto por softwares de tradução assistida por computador, ou produzir o texto primeiramente em um idioma do qual não se tem domínio e depois traduzi-lo automaticamente para seu próprio idioma materno podem, em alguns casos, prejudicar o trabalho de investigação linguística. Mas vale se considerar que mesmo redigindo um texto em um idioma o qual não se fala fluentemente e mesmo pós-processando textos em softwares de tradução, ainda assim aquelas evidências conterão dados únicos do idioleto daquele falante.

Ataques perpetrados por grupos hackers pluricêntricos podem apresentar pedidos de resgate, manifestos e outros elementos passíveis de análise linguística que tragam evidências de que tenham sido produzidos por falantes que dominem com fluência ou a nível de língua materna mais de um idioma, o que pode indicar que a composição destes grupos possui células presentes em mais de um país.

No entanto, quando faltam as evidências linguísticas diretas presentes em ciberataques, em se tratando especificamente de amostras como pedidos de resgates, manifestos e outros textos, é possível, ainda, proceder com outro tipo de análise linguística: a da estruturação lógica utilizada no código-fonte. Cada programador possui seu *footprint* que é definido pela forma como organiza seu próprio algoritmo e, por isso, os códigos-fontes são também importantes evidências que permitem indicar idade, origem e outras informações dos indivíduos que se ocultam por trás dos códigos maliciosos.

Por fim, por mais que se tente produzir o crime perfeito, as ciências forenses tornam-se dia após dia cada vez mais sofisticadas, e, no que se relaciona com a Linguística Forense como ferramenta para a Inteligência Cibernética, as características próprias das linguagens, sejam elas o idioma falado nativamente ou uma linguagem de programação, tornam muito difíceis a possibilidade de inserção de falsas pistas, sem que outras pistas surjam a partir daí.

* Prof. Dr. Leonardo Perin Vichi
Doutor em História Social pela UFRJ/Freie Universität Berlin
Pesquisador na Escola de Comando e Estado-Maior do Exército - ECEME
contato@leonardovichi.com
Vinculação ao NEEDS: Mar/2019