Datenschutzkonforme Anonymisierung von Datenverkehr auf einem Vermittlungssystem

Thomas Gamer, Christoph Mayer, Marcus Schöller

Institut für Telematik
Universität Karlsruhe (TH)

Aufzeichnungen von Datenverkehr – so genannte Netzwerktraces – werden in hohem Maße für Forschung und Industrie benötigt. Mit Hilfe von mitgeschnittenem Datenverkehr ist es möglich, Eigenschaften und Verhalten von Netzen sowie von deren Komponenten, Diensten und Benutzern zu analysieren. Auch die Entwicklung spezieller Anwendungen – wie beispielsweise eines Systems zur Angriffserkennung – benötigt Datenverkehr zur Analyse und Evaluierung des konzipierten Systems. Die Unvorhersehbarkeit von Nutzerverhalten, nicht protokollkonforme Pakete, Angriffe und die allgemein sehr große Diversität in realen Netzdaten legen hierbei deren Einsatz nahe. Künstlich erzeugter Verkehr hingegen kann weder im Hinblick auf Benutzerverhalten noch auf Verkehrscharakteristiken die Realität nachbilden. Reale Daten garantieren daher, dass neu entwickelte Systeme unter realistischen und nicht an bestimmte Tests angepassten Bedingungen getestet werden können bevor sie tatsächlich eingesetzt werden.

Die Aufzeichnung von realem Datenverkehr wird durch datenschutzrechtliche Bestimmungen eingeschränkt, d.h. Netzwerktraces müssen bei Aufzeichnung und Weitergabe einer Anonymisierung unterzogen werden. Diese Anonymisierung muss sicher, schnell, robust und leicht auf neue Protokolle und Anonymisierungsprimitive erweiterbar sein. Zur Durchführung einer solchen Anonymisierung wurde am Institut für Telematik der Universität Karlsruhe (TH) ein flexibles und leicht erweiterbares Framework zur datenschutzkonformen Anonymisierung von Datenverkehr auf einem Vermittlungssystem entwickelt.

Bei der Aufzeichnung und Weitergabe von Netzdaten schreiben die datenschutzrechtlichen Bestimmungen lediglich eine Anonymisierung vor, welche die Zuordnung einzelner Pakete zur Identität eines Nutzers unmöglich macht sowie personenbezogene Daten eines Nutzers schützt. Die Anonymisierung von Datenverkehr steht jedoch zusätzlich vor der Herausforderung, einen Kompromiss zwischen Sicherheit und Nutzen der anonymisierten Daten zu erzielen. Dieser Kompromiss hängt direkt vom Vertrauensverhältnis zwischen der Partei, welche die Daten aufzeichnet, und der Partei, an welche die Daten weitergegeben werden, ab. Eine Veröffentlichung der aufgezeichneten Netzdaten im Internet erfordert beispielsweise eine wesentlich höhere Sicherheit – d.h. es müssen wesentlich mehr sensible Daten anonymisiert werden – als eine unternehmensinterne Weitergabe. Diese erhöhte Sicherheit wird zwar vom Gesetzgeber nicht gefordert, ist jedoch in hohem Maße für das Unternehmen selbst von Interesse. Durch eine sicherere Anonymisierung sinkt allerdings der Nutzen von Netzwerktraces, da dadurch beispielsweise statistische Daten des aufgezeichneten Verkehrs verloren gehen.

Die Hauptrisiken bei einer Veröffentlichung von Netzdaten im Internet bestehen darin, sensible Nutzerdaten und Informationen über das Netzwerk selbst preiszugeben. Auf Anwendungsschicht beinhalten die Nutzdaten eines Pakets beispielsweise Informationen aus Datenbankabfragen oder Emailkommunikation. Des Weiteren können durch Analyse der Header tieferer Schichten z.B. Informationen über die Netzstruktur mit all ihren Diensten und Komponenten gewonnen werden. Eine „unsichere“ Anonymisierung ermöglicht Angreifern daher, potentielle Angriffsziele und Schwachstellen im Netzwerk zu identifizieren.

Das von uns entwickelte Framework *pktanon* beruht auf folgenden Eigenschaften: lose Kopplung der Protokollschichten, defensive Anonymisierung, Wohlgeformtheit der anonymisierten Daten, hohe Verarbeitungsgeschwindigkeiten und leichte Erweiterbarkeit. Um

beliebige Protokollschachtelungen verarbeiten zu können, werden die einzelnen Protokollschichten eines Pakets unabhängig voneinander behandelt, d.h. sie sind nur lose miteinander verbunden. Dadurch ist es beispielsweise möglich, völlig transparent für die Protokoll-spezifischen Klassen, auch IP-in-IP-Pakete oder ICMP-Pakete, die wiederum IP-Pakete enthalten, zu unterstützen. Das Prinzip der defensiven Anonymisierung basiert auf der Idee, für jedes zu bearbeitende Paket ein neues, leeres Paket zu erstellen und dieses nach und nach mit den anonymisierten Protokollschichten des originalen Pakets zu füllen. Durch die Erstellung eines neuen Pakets müssen alle Felder einer Protokollschicht explizit anonymisiert oder kopiert werden, um ein hohes Maß an Sicherheit zu gewährleisten und die Veröffentlichung sensibler Daten zu vermeiden. Würden bei der Bearbeitung eines Pakets die Informationen direkt in den einzelnen Protokollschichten des originalen Pakets anonymisiert, kann dies zu Problemen führen, falls sensible Daten bei der Anonymisierung übersehen werden. Des Weiteren muss jede Protokoll-spezifische Klasse für die Wohlgeformtheit der anonymisierten Protokollschicht sorgen, indem z.B. Prüfsummen und Längenfelder neu berechnet werden, falls Änderungen im Vergleich zum originalen Paket vorgenommen wurden. Dadurch ist gewährleistet, dass Netzwerktraces nach der Anonymisierung wohlgeformt sind und ohne Probleme weiterverarbeitet werden können.

Hohe Verarbeitungsgeschwindigkeiten und ein sorgfältiger Umgang mit Ressourcen sind bei der Anonymisierung aus mehreren Gründen notwendig: zum einen müssen große Datenmengen verarbeitet werden, was unter Umständen sehr viel Zeit und Rechenleistung in Anspruch nehmen kann. Zum anderen ist es aus Sicht des Datenschutzes notwendig, Netzdaten online zu anonymisieren, d.h. die Netzdaten werden nicht erst lokal gespeichert und dann einer Anonymisierung unterzogen, sondern sie werden on-the-fly anonymisiert und somit nur in datenschutzkonformer Art auf dem Hintergrundspeicher abgelegt. Dies erleichtert die Aufzeichnung der Daten und bietet weniger Angriffsmöglichkeiten.

Das Framework bietet außerdem bereits eine Vielzahl einsetzbarer Anonymisierungsprimitive an und ermöglicht dadurch eine einfache Erweiterbarkeit um neue Protokoll-spezifische Klassen. Neu einzufügende Klassen müssen sich nur um den Aufbau ihres eigenen Protokolls kümmern und spezifizieren, welche Felder bei der Bearbeitung kopiert bzw. anonymisiert werden sollen. Für jedes zu anonymisierende Feld kann unabhängig von anderen Feldern gewählt werden, welches Anonymisierungsprimitiv verwendet werden soll. Durch die lose Kopplung der Protokollschichten und die Bereitstellung der Hauptfunktionalität durch das Framework selbst können Protokoll-spezifische Klassen daher einfach ausgetauscht, deaktiviert oder neu hinzugefügt werden.

Sicherheit und Erweiterbarkeit sind die Hauptkriterien für ein Anonymisierungstool. Diese werden in *pktanon* durch defensive Anonymisierung und Verlagerung der Hauptfunktionalität in das Framework erfüllt. Das Framework erreicht Verarbeitungsgeschwindigkeiten von etwa 100 MBit/s und kann damit für Live-Anonymisierungen verwendet werden. Durch das Zusammenspiel mit Programmen wie *tcpdump* ist es somit möglich, Netzdaten on-the-fly aufzuzeichnen und anonymisiert auf die Festplatte zu schreiben bzw. weiterzuverarbeiten. Außerdem können neue Anonymisierungsprimitive einfach hinzugefügt werden.

Um den Einsatz in unterschiedlichen Umgebungen – d.h. unterschiedliche Anforderungen an die Sicherheit der Anonymisierung – zu vereinfachen, soll das bestehende Framework derart weiterentwickelt werden, dass XML-basierte Konfigurationsprofile verwendet werden können. Dadurch ist es möglich, vordefinierte Konfigurationen zu erstellen, welche unterschiedliche Vertrauensverhältnisse modellieren. Hierbei können beispielsweise auch unterschiedliche juristische Anforderungen einfließen, die es in diesem Bereich nicht versierten Benutzern ermöglichen, die benötigte Sicherheit und juristische Absicherung der Anonymisierung sicherzustellen. Ein weiteres Ziel zukünftiger Weiterentwicklungen ist die Unterstützung von Geschwindigkeiten im Gigabit-Bereich, um Anonymisierungen on-the-fly auch im Netzinneren des Internets zu ermöglichen. Dies ist vor allem durch den im Vergleich zum Randbereich des Internets häufig noch ungefilterten Verkehr interessant.