วันที่ 16 มิถุนายน 2569

Agentjacking Attack หลอก AI Coding Agent ให้รันโค้ดอันตรายบนเครื่องนักพัฒนา

นักวิจัยด้านความปลอดภัยไซเบอร์ได้เปิดเผยการโจมตีรูปแบบใหม่ที่สามารถหลอกให้AI Coding Agent รันโค้ดตามที่ผู้โจมตีต้องการบนเครื่องของนักพัฒนาได้ การโจมตีนี้มีชื่อว่า Agentjacking ซึ่งถูกค้นพบโดย Tenet Security โดยสามารถเริ่มต้นการโจมตีได้ผ่านรายงานข้อผิดพลาดปลอมที่ถูกสร้างขึ้นโดยใช้Sentry ซึ่งเป็นแพลตฟอร์มโอเพนซอร์สสำหรับติดตามข้อผิดพลาดและตรวจสอบประสิทธิภาพของระบบ

การโจมตีนี้อาศัยช่องโหว่ด้านสถาปัตยกรรมที่สำคัญ ซึ่งเกิดขึ้นระหว่างระบบรับข้อมูลเหตุการณ์ของ Sentry (ที่ยอมรับข้อมูลใดๆ จากผู้ที่มี DSN) และ Sentry MCP Server (ที่ส่งข้อมูลเหล่านั้นกลับไปให้ AI Agent ในรูปแบบที่ถูกมองว่าเป็นข้อมูลที่เชื่อถือได้) นักวิจัยด้านความปลอดภัย Ron Bobrov, Barak Sternberg และ Nevo Poran กล่าว

กลไกการโจมตีและผลกระทบที่อาจเกิดขึ้น

แนวคิดของการโจมตีคือการแทรกข้อมูลที่ถูกสร้างขึ้นมาโดยเฉพาะเข้าไปในเหตุการณ์ข้อผิดพลาดของSentry จากนั้น AI Coding Agent เช่น Claude Code และ Cursor จะตีความข้อมูลดังกล่าวว่าเป็นขั้นตอนการแก้ไขปัญหาที่ถูกต้อง และดำเนินการรันโค้ดที่ผู้โจมตีควบคุมไว้

หากการโจมตีสำเร็จ อาจนำไปสู่การเปิดเผยข้อมูลสำคัญ เช่น ตัวแปรสภาพแวดล้อม (Environment Variables), ข้อมูลรับรอง Git, URL ของ Private Repository รวมถึงข้อมูลระบุตัวตนของนักพัฒนา โดยไม่จำเป็นต้องใช้วิธีฟิชชิงหรือเจาะระบบเซิร์ฟเวอร์ล่วงหน้าแต่อย่างใด สาเหตุของปัญหานี้มาจากความเชื่อมั่นโดยปริยายที่เกิดขึ้นเมื่อมีการเชื่อมต่อกับบริการภายนอกผ่านModel Context Protocol (MCP) เนื่องจาก AI Agent ไม่สามารถแยกแยะได้ว่าเหตุการณ์ข้อผิดพลาดนั้นเกิดจากแอปพลิเคชันล่มจริง หรือเป็นข้อมูลที่ผู้โจมตีแทรกเข้ามา จึงทำให้เกิดช่องทางสำหรับการรันโค้ดตามที่ผู้โจมตีต้องการเมื่อ Agent ประมวลผลข้อมูลดังกล่าว

ลำดับขั้นตอนการโจมตีที่ Tenet พัฒนาขึ้น

ลำดับการโจมตีที่ Tenet พัฒนาขึ้นมีดังนี้:

- ผู้โจมตีค้นหา Sentry Data Source Name (DSN) ของเป้าหมาย ซึ่งเป็นข้อมูลสาธารณะที่ใช้สำหรับส่งข้อมูลเข้า และมักฝังอยู่ในเว็บไซต์
- ผู้โจมตีส่งเหตุการณ์ข้อผิดพลาดที่เป็นอันตรายไปยัง Sentry Ingest Endpoint ผ่านคำขอแบบ POST โดยใช้ DSN ดังกล่าว
- เหตุการณ์ที่ถูกแทรกจะมี Markdown ที่ถูกจัดรูปแบบอย่างพิถีพิถันอยู่ภายในช่อง Message และชื่อของ Context Keys เมื่อ Sentry MCP Server ส่งข้อมูลนี้กลับไปยัง AI Agent ข้อมูลจะถูกแสดงผลในรูปแบบที่มีโครงสร้างและมีหน้าตาเหมือนกับ Template ของระบบ Sentry อย่างสมบูรณ์
- เมื่อนักพัฒนาสั่ง AI Coding Agent ด้วยคำสั่งลักษณะเช่น "แก้ไขปัญหา Sentry ที่ยังไม่ได้รับการแก้ไข" หรือคำสั่งในลักษณะเดียวกัน Agent จะเชื่อมต่อไปยัง Sentry ผ่าน MCP และได้รับเหตุการณ์ที่เป็นอันตรายดังกล่าว
- Agent จะรันโค้ดอันตราย ซึ่งทำงานภายใต้สิทธิ์ทั้งหมดของนักพัฒนา

จุดเด่นและความแนบเนียนของการโจมตี

ผู้โจมตีไม่จำเป็นต้องแตะต้องโครงสร้างพื้นฐานของเหยื่อเลยแม้แต่น้อย นักวิจัยอธิบายคำสั่งอันตรายถูกส่งมาในรูปแบบของ Resolution ที่ดูเหมือนเป็นคำแนะนำปกติภายในรายงานข้อผิดพลาดทั่วไป เมื่อผู้พัฒนาขอให้ AI Agent ช่วยแก้ปัญหา Sentry Agent จะมองว่าคำสั่งของผู้โจมตีเป็นคำแนะนำที่เชื่อถือได้ และดำเนินการรันคำสั่งนั้นด้วยสิทธิ์ของนักพัฒนาเอง บนเครื่องของนักพัฒนาเอง

Agentjacking มีความโดดเด่นตรงที่มุ่งเป้าไปยัง AI Agent ที่นักพัฒนาไว้วางใจและใช้งานอยู่เป็นประจำ โดยใช้ Sentry DSN เป็นจุดเริ่มต้นของการโจมตี นอกจากนี้ การแทรก Markdown ยังถูกออกแบบให้แสดงผลในลักษณะที่ Agent ไม่สามารถแยกความแตกต่างระหว่างข้อมูลที่ถูกต้องกับข้อมูลที่ผู้โจมตีสร้างขึ้นได้

สถิติความสำเร็จและการตอบสนองจาก Sentry

บริษัทด้านความปลอดภัย AI ระบุว่าพบองค์กรอย่างน้อย2,388 แห่งที่มี DSN ซึ่งสามารถถูกใช้ในการโจมตีได้ และจากการทดสอบในสภาพแวดล้อมควบคุมกับองค์กรมากกว่า 100 แห่ง พบว่าอัตราความสำเร็จของการโจมตีผ่านเหตุการณ์ที่ถูกแทรกข้อมูลอยู่ที่ 85% ครอบคลุม AI Coding Assistant ที่ได้รับความนิยมมากที่สุดหลายรายการ

ในส่วนของ Sentry ได้รับทราบถึงปัญหาดังกล่าวแล้ว แต่เลือกที่จะไม่แก้ไขโดยตรง เนื่องจากมองว่า "ไม่สามารถป้องกันได้ในทางเทคนิค" อย่างไรก็ตาม บริษัทได้เปิดใช้งานตัวกรองเนื้อหาระดับโลกเพื่อบล็อก "Payload Stringรูปแบบเฉพาะ" ที่ใช้ในการโจมตี

บทสรุป: ภัยคุกคามที่สามารถหลบหลีกระบบความปลอดภัย

"ในขณะที่องค์กรต่างๆ เร่งนำ AI Coding Agent มาใช้งาน งานวิจัยนี้แสดงให้เห็นว่า Agent เองได้กลายเป็นพื้นผิวการโจมตีรูปแบบใหม่ และสามารถถูกใช้ย้อนกลับมาโจมตีนักพัฒนาที่ไว้วางใจมันได้ โดยอาศัยเพียงข้อมูลที่องค์กรเผยแพร่สู่สาธารณะด้วยตัวเอง" Tenet กล่าว

การโจมตีนี้สามารถหลีกเลี่ยง EDR, WAF, IAM, VPN, Cloudflare และไฟร์วอลล์ได้ เนื่องจากไม่มีพฤติกรรมใดที่ถูกมองว่าเป็นอันตรายให้ตรวจจับ ทุกขั้นตอนที่เกิดขึ้นในกระบวนการล้วนเป็นการกระทำที่ได้รับอนุญาตอย่างถูกต้องทั้งสิ้น

ข้อมูลอ้างอิง

Jun 12, 2026, By Ravie Lakshmanan

- https://thehackernews.com/2026/06/agentjacking-attack-tricks-ai-coding.html