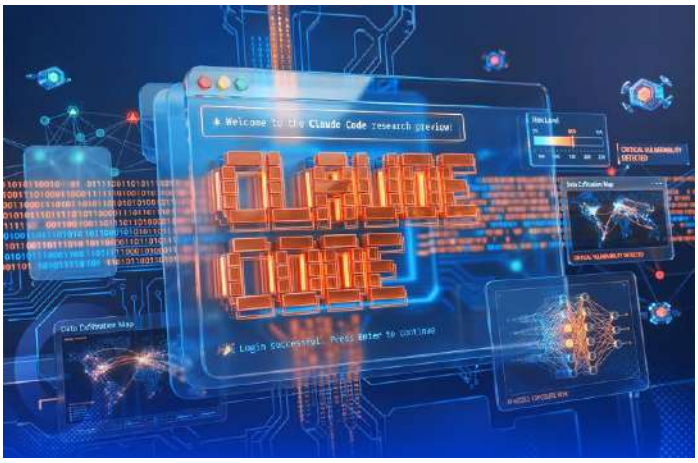
วันที่ 2 เมษายน 2569

ซอร์สโค้ด Claude Code รั่วไหล หลังพบความผิดพลาดในการอัปเดตแพ็กเกจ npm



บริษัท Anthropic ยืนยันเมื่อวันอังคารที่ผ่านมาว่า ซอร์สโค้ดภายในของผู้ช่วยเขียนโค้ดด้วย AI ยอดนิยมอย่าง Claude Code ถูกเผยแพร่ออกมาโดยไม่ได้ตั้งใจ เนื่องจากความผิดพลาดของมนุษย์ระหว่างขั้นตอนการแพ็กเกจไฟล์ โฆษกของ Anthropic กล่าวในแถลงการณ์ที่ส่งให้กับ CNBC News ว่า “ไม่มีข้อมูลลูกค้าหรือข้อมูลรับรองความปลอดภัยใดๆ ที่เกี่ยวข้องหรือถูกเปิดเผย เหตุการณ์นี้เกิดจากปัญหาในการแพ็กเกจไฟล์ที่เกิดจากความผิดพลาดของมนุษย์ ไม่ใช่การเจาะระบบด้านความปลอดภัย และเรากำลังเพิ่มมาตรการเพื่อป้องกันไม่ให้เกิดเหตุการณ์แบบนี้อีก”

การแพร่กระจายของข้อมูลความลับ

เหตุการณ์นี้ถูกค้นพบหลังจากบริษัทปล่อย Claude Code เวอร์ชัน 2.1.88 บน npm โดยผู้ใช้พบว่าแพ็กเกจดังกล่าวมี source map file อยู่ภายใน ซึ่งสามารถใช้เข้าถึงซอร์สโค้ดของ Claude Code ได้ ซอร์สโค้ดที่หลุดออกมามีขนาดใหญ่ ประกอบด้วยไฟล์ TypeScript เกือบ 2,000 ไฟล์ และมากกว่า 512,000 บรรทัดของโค้ด ปัจจุบันเวอร์ชันดังกล่าวถูกลบออกจาก npm แล้ว และไม่สามารถดาวน์โหลดได้อีก

นักวิจัยด้านความปลอดภัยชื่อ Chaofan Shou เป็นบุคคลแรกที่ออกมาเปิดเผยเรื่องนี้ผ่านแพลตฟอร์ม X โดยระบุว่า “Claude code source code has been leaked via a map file in their npm registry!”หรือก็คือ“ซอร์สโค้ดของ Claude ถูกเปิดเผย (หลุด) ผ่านไฟล์ map ใน npm registry ของพวกเขา”

ซึ่งโพสต์ดังกล่าวมียอดเข้าชมมากกว่า 28.8 ล้านครั้ง แล้ว ซ้ำร้าย ซอร์สโค้ดดังกล่าวยังถูกนำไปเผยแพร่ต่อบน GitHub repository แบบสาธารณะ โดยมีผู้กดดาว มากกว่า 84,000 ครั้ง และมีการ Fork ไปแล้วกว่า 82,000 ครั้ง การหลุดของข้อมูลระดับนี้ส่งผลให้นักพัฒนาทั่วไปและบริษัทคู่แข่งสามารถเจาะลึกโครงสร้างการทำงานของเครื่องมือเขียนโค้ดยอดนิยมตัวนี้ได้อย่างทะลุปรุโปร่ง

เจาะลึกฟีเจอร์ที่ถูกซ่อนไว้

ผู้ใช้ที่เข้าไปวิเคราะห์โค้ดได้เผยรายละเอียดเกี่ยวกับ ระบบหน่วยความจำแบบ self-healing ของ Claude Code ซึ่งถูกออกแบบมาเพื่อแก้ข้อจำกัดเรื่อง context window ของโมเดล รวมถึงส่วนประกอบภายในอื่นๆ

ตัวอย่างระบบภายในที่ถูกพบ ได้แก่

- ระบบเครื่องมือ (tools system) ที่ช่วยให้ทำงานต่าง ๆ เช่น อ่านไฟล์ หรือรันคำสั่ง bash
- query engine สำหรับจัดการการเรียกใช้ API ของโมเดลภาษา (LLM)
- ระบบ multi-agent orchestration ที่สามารถสร้าง “sub-agents” หรือกลุ่มเอเจนต์ย่อยเพื่อทำงานที่ซับซ้อน
- ชั้นการสื่อสารสองทาง (bidirectional communication layer) ที่เชื่อมต่อส่วนขยายของ IDE กับ CLI ของ Claude Code

นอกจากโครงสร้างพื้นฐาน โค้ดยังเผยให้เห็นฟีเจอร์หนึ่งชื่อ "KAIROS" ซึ่งเปลี่ยน Claude สามารถทำงานเป็นเอเจนต์เบื้องหลังที่ทำงานต่อเนื่องเช่น คอยตรวจและแก้ข้อผิดพลาดเป็นระยะ หรือรันงานบางอย่างเองโดยไม่ต้องรอคำสั่งจากผู้ใช้ และยังสามารถส่ง push notification ไปยังผู้ใช้ได้ด้วย นอกจากนี้ยังมีโหมดใหม่ชื่อ “dream mode” ซึ่งออกแบบมาให้ Claude สามารถ “คิด” อยู่เบื้องหลังตลอดเวลา เพื่อพัฒนาไอเดียหรือปรับปรุงแนวคิดเดิมอย่างต่อเนื่อง และที่น่าจับตามองคือ "Undercover Mode" หรือโหมดแฝงตัว ที่ตั้งค่าให้ Claude ซึ่งถูกออกแบบมาเพื่อให้ Claude สามารถมีส่วนร่วมกับโปรเจกต์โอเพ่นซอร์สแบบไม่เปิดเผยตัว ใน system prompt มีข้อความระบุว่า “You are operating UNDERCOVER in a PUBLIC/OPEN-SOURCE repository. Your commit messages, PR titles, and PR bodies MUST NOT contain ANY Anthropic-internal information. Do not blow your cover.”

นอกจากนี้ยังพบความพยายามของ Anthropic ในการป้องกันการโจมตีแบบ model distillation โดยระบบมีการฝังเครื่องมือปลอม (fake tool definitions) ลงในคำขอ API เพื่อทำให้ข้อมูลที่ถูกดึงไปใช้ฝึกโมเดลของคู่แข่งมีข้อมูลที่ปนเปื้อน

ผลกระทบและภัยคุกคามทางไซเบอร์

บริษัทความปลอดภัย AI Straiker ชี้ให้เห็นว่า แทนที่ผู้โจมตีจะต้องลองสุ่มวิธี jailbreak หรือ prompt injection เพราะพวกเขาสามารถศึกษาได้โดยตรงว่าข้อมูลไหลผ่านระบบจัดการ context แบบ 4 ขั้นตอนของ Claude Code อย่างไร และสามารถออกแบบ payload ที่ยังคงอยู่ในระบบได้ตลอดเซสชัน

สิ่งที่น่ากังวลมากกว่านั้นคือผลกระทบจาก การโจมตีซัพพลายเชน Axios ผู้ใช้ที่ติดตั้งหรืออัปเดต Claude Code ผ่าน npm ในวันที่ 31 มีนาคม 2026 ระหว่างเวลา 00:21 ถึง 03:29 UTC อาจได้รับเวอร์ชันของ HTTP client ที่ถูกฝังมัลแวร์ไว้ ซึ่งเป็นโทรจันควบคุมเครื่องจากระยะไกลที่ทำงานได้หลายแพลตฟอร์ม ผู้ใช้ถูกแนะนำให้ ลดเวอร์ชันลงทันทีไปยังเวอร์ชันที่ปลอดภัย และเปลี่ยนข้อมูลลับทั้งหมด

ระวังโดเมนปลอมดักข้อมูล

ผู้โจมตียังเริ่มใช้โอกาสจากเหตุการณ์นี้ในการ typosquat ชื่อแพ็กเกจ npm ภายใน เพื่อโจมตีนักพัฒนาที่พยายามคอมไพล์ซอร์สโค้ด Claude Code ที่หลุดออกมา

แพ็กเกจเหล่านี้ถูกเผยแพร่โดยผู้ใช้ชื่อ “pacifier136” ได้แก่

- audio-capture-napi
- color-diff-napi
- image-processor-napi
- modifiers-napi
- url-handler-napi

นักวิจัยด้านความปลอดภัย Clément Dumas กล่าวบน X ว่า “ตอนนี้มันยังเป็นเพียงสตับว่างๆ (module.exports = {}) แต่การโจมตีแบบนี้มักทำแบบนี้ก่อน คือจองชื่อแพ็กเกจไว้ รอให้มีคนดาวน์โหลดจำนวนมาก แล้วจึงปล่อยอัปเดตที่เป็นมัลแวร์ในภายหลัง” เหตุการณ์นี้ถือเป็นความผิดพลาดครั้งใหญ่ครั้งที่สองของ Anthropic ภายในหนึ่งสัปดาห์ หลังจากที่ก่อนหน้านี้ข้อมูลเกี่ยวกับโมเดล AI ตัวใหม่ของบริษัท พร้อมข้อมูลภายในบางส่วน ถูกเปิดให้เข้าถึงได้ผ่านระบบ Content Management System (CMS) ของบริษัทโดยไม่ได้ตั้งใจ ต่อมา Anthropic ยอมรับว่ากำลังทดสอบโมเดลดังกล่าวกับลูกค้าที่ได้รับสิทธิ์เข้าถึงก่อน โดยระบุว่าเป็นโมเดลที่ “มีความสามารถมากที่สุดเท่าที่บริษัทเคยสร้างมา” ตามรายงานของ Fortune

ข้อมูลอ้างอิง

Apr 1, 2026, By Ravie Lakshmanan

- https://thehackernews.com/2026/04/claude-code-tleaked-via-npm-packaging.html