Talking Past Each Other: The Legal and Technical Challenges of Harmful Web Scraping

White Paper Fast Facts

I. Web scraping – the good, the bad, the ugly

- **What is Web Scraping?**: Web scraping is the use of automated tools, such as bots, to collect data in bulk from websites. Many website owners describe scraping that violates a website’s terms of service (TOS) as “unauthorized web scraping.”
- **Terms of Service**: Terms of service and other policies protect the interests of website owners and define some level of protection for users. By contrast, scrapers do not face such consequences when they misuse data. Scrapers have no relationship with the users whose data they have scraped and have made no promises to safeguard their data.
- **Risks of Scraping**: Unwanted web scraping exploits personal information collected for entirely different purposes, violates personal privacy, and poses cybersecurity risks. Scammers and other threat actors liberally use bots, including web scrapers, to harvest sensitive information.
- **Privacy-Invasive Web Scraping**: Some notable unauthorized web scraping examples include Clearview AI scraping billions of photographs for facial recognition and HiQ Labs scraping LinkedIn profiles for employer surveillance. Dating websites are often targeted for harmful web scraping, leading to privacy breaches and potential blackmail.

II. Cybersecurity, web scraping and the “three A’s”: authentication, authorization, access control

- Aggressive and unwanted scraping often violate basic security principles, bypassing authentication, authorization, and access controls that website owners use to safeguard their systems.
- Authentication and authorization are distinct concepts.
- Authentication is not synonymous with logins and passwords and not all forms of authentication are based on credentials or establishing a user’s identity.
- Much of the open web requires neither authentication, nor authorization.
- Though usually coupled, authentication is not a necessary prerequisite for authorization, which can occur before, after, or even without authentication.
- Website owners who wish to discourage scrapers can ban scraping in their TOS, discourage web crawlers from some portions of their sites in their “robots.txt” files, or take legal steps.
- Website owners also have technical anti-scraping countermeasures including rate-limiting log-ins, blocking IP addresses, using CAPTCHAs, and dynamically changing elements of the website’s HTML code, which can all meet the technical definition of authentication.

III. Scraping and the Computer Fraud and Abuse Act (CFAA): why the courts are confused

- Federal courts have been confused for years about how to apply the main federal anti-hacking statute – the Computer Fraud and Abuse Act (CFAA) – to unwanted computer behaviors.
- The misunderstanding is due to the outdated wording of the CFAA, no clear definition of authorization, omission of the concept of authentication, and the frequent misinterpretation of technical terms used in the CFAA to refer to legal concepts rather than technical ones.

---

1 [White Paper](#) by Timothy H. Edgar
In *Van Buren v. United States*, the Supreme Court resolved the debate by adopting a narrower interpretation of authorization consistent with the technical concept of authorization instead of the legal one, though left unanswered questions regarding authorization "gate" requirements.

The Ninth Circuit's decision in *LinkedIn v. HiQ Labs*, which interpreted the CFAA as a trespass statute and ruled against LinkedIn despite anti-scraping and anti-bot technologies in place demonstrated that misunderstanding of the CFAA continues.

Because the United States lacks a comprehensive privacy law, a weakening of existing legal protections against unwanted scraping of personal data would mark another step in the journey towards a "wild west" for exploitation of such data.

**IV. Legislative Solutions for Harmful Scraping**

- Addressing the social harms of harmful scraping requires Congress to act. Congress could do so by amending the CFAA, by addressing the problem of harmful scraping in comprehensive federal privacy legislation, or by doing both.
- Personal data scraped from EU subjects is protected under the GDPR, regardless of its public availability.
- The proposed American Data Privacy and Protection Act (ADPPA), which aims to reconcile the interests of industry and privacy advocates, does not protect personal data that is publicly available.
- Any legislation that seeks to address the problems of harmful web scraping will have to address the global nature of the problem and empower website owners to deploy technical countermeasures.
- The demand for scraped data has become even greater with the increasing sophistication of new digital services based on Artificial Intelligence/Machine Learning (AI). Proposed ethical and regulatory frameworks for AI should take into account the ethical and regulatory problems of using scraped personal data to train ML algorithms.
- The demand for scraped data is increasing even as the norms, laws, and regulations that have limited such scraping have eroded.

**V. Conclusion**

- Harmful web scraping is a cybersecurity threat that demands attention due to increasing demand for scraped data and weakening norms, laws, and regulations.
- Incentives should be provided to website owners to limit harmful scraping and protect user privacy and security.
- Policymakers should take steps to prevent harmful scraping, while ensuring appropriate exceptions for scraping for valid commercial purposes and for legitimate and ethical research.
- Collaboration between lawyers, policymakers, and computer security experts is critical to addressing the problem of harmful web scraping.
- Only comprehensive legislation that protects privacy and personal data can offer data subjects effective remedies against the misuse of personal data.
- Solutions to harmful web scraping should transcend jurisdictions and legal systems and include technical countermeasures and properly-informed laws.