Crowdsourcing and Snowball Sampling: Addressing Challenges in Data Collection from Social Media and News Platforms

Iain J. Cruickshank,[1] **Ian Kloo** [1]

[1] United States Military Academy, West Point, NY, USA
iain.cruickshank@westpoint.edu, ian.kloo@westpoint.edu

Abstract

The social media landscape is becoming increasingly fragmented and restrictive for researchers. Specifically, fewer and fewer social media sites allow programmatic or API access to their data, even for research purposes. In this paper, we propose a two-stage methodology to overcome challenges in accessing data from social media and news websites for computational social science research. Firstly, we leverage crowdsourcing to gather relevant links from users who frequently share links to posts or websites, including those within walled gardens on social media platforms. Subsequently, we employ snowball sampling to expand the dataset from the collected links, effectively identifying valuable discussions within social media platforms. Despite technical hurdles such as scraping data from sources lacking APIs and the difficulty in finding reliable seed links, our approach offers a systematic means of gathering pertinent data for computational social science research. It is hoped this extended abstract detailing past ways of collecting data can serve as a discussion point for envisioning new methods of collecting data for research.

Amidst the evolving landscape of social media and online news platforms, researchers face mounting obstacles in accessing data critical for computational social science inquiries. From strict API restrictions on popular social networks to the labyrinthine structure of modern news websites, navigating these digital terrains requires innovative strategies to ensure effective data collection and analysis. Firstly, several popular social media sites now impose restrictions on access to their APIs, citing various reasons, thus hindering data collection efforts by researchers (Axel Bruns and Bruns 2019; Balakrishnan et al. 2022; Durbin 2023). In addition, the social media landscape is becoming more fragmented. While platforms like Snapchat, Discord, or Telegram offer public discussion forums akin to traditional social media, accessing data from these forums requires specific links to the respective discussion forums (Kloo and Carley 2023). These 'walled gardens' inherent in social media platforms pose significant hurdles to data collection, as they cannot be simply searched by keywords (Kloo and Carley 2023).

Additionally, it is crucial to note that collecting data from regular websites, such as news websites, is equally vital for computational social science research and presents its own set of challenges. Scraping news websites, similar to navigating walled gardens in social media platforms, requires knowledge of specific URLs; attempting to scrape entire sites often results in rate-limiting issues and may yield irrelevant data (Cruickshank and Carley 2020). With the vast number of online news sites and the proliferation of 'pink slime' news sources, identifying relevant links becomes arduous (Lepird and Carley 2023b,a). Therefore, researchers require a systematic approach to gather pertinent links to access and leverage data both for walled gardens in social media sites and to collect relevant news articles.

Methdology

In order to address the challenges associated with data collection from walled gardens in social media sites and webpages on news sites, we propose the use of a two-stage methodology. This two-stage methodology, depicted in Figure 1, consists of crowdsourcing links for pages or subdomains and then snowball sampling based on the data collected from the crowd-sourced links. For the first stage, previous work has found that users, particularly social media users on microblogging sites that restrict the amount of text that a user can post, frequently share links to other social media posts or websites (Cruickshank et al. 2021; Ginossar et al. 2022; Ng, Cruickshank, and Carley 2022; Cruickshank and Carley 2020). This link-sharing behavior often includes links to other social media sites, including links to posts in walled gardens on social media sites (Ng, Cruickshank, and Carley 2022; Ginossar et al. 2022; Kloo and Carley 2023). This link-sharing behavior can occur for a number of reasons, but it is often done to provide further explanation or evidence for claims in a user's post when done on a microblogging social media site (Cruickshank et al. 2021; Ginossar et al. 2022). As such, these links are often relevant for behavioral and other social science research. Furthermore, disinformation actors frequently use social media sites to spread links to disinformation websites or particular walled gardens, which often contain valuable data for computational social science researchers (Ng, Cruickshank, and Carley 2022; Ng and Taeihagh 2021; Balakrishnan et al. 2022). In addition, crowd-sourcing on social media has proven useful for finding valuable information in other contexts, such as disaster relief (Gao, Barbier, and Goolsby 2011). Thus,

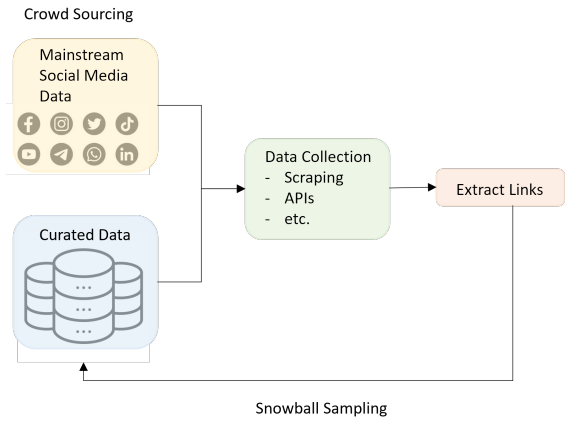

Figure 1: Using an iterative process of crowdsourcing and snowball sampling to find data in walled gardens and news sites for collection

the first step in overcoming the 'walls' of walled gardens is to leverage crowdsourcing to find relevant links to things like posts in walled gardens of social media sites or particular news stories, which overcomes the inability to widely search those social media or news sites.

Once we have found some seed links, the next step is to snowball sample out from those links. Snowball sampling, or link tracing, has been frequently used on social networks as a means of building usable samples of social media data (Leighton et al. 2021; Heckathorn and Cameron 2017; Chan 2020). The process of snowball sampling involves extracting any additional links present in the data collected from the seed links, which were identified in the previous step of crowdsourcing, and then collecting the data from those links as well. This process can be iterative as resources allow. Previous works have found this process of snowball sampling with links present in social media data is an effective way of identifying all relevant walled garden discussions within a social media platform that has walled gardens, as users often share links to related discussions in other walled gardens. For example, in (Kloo and Carley 2023), the authors used links present in seed Telegram discussions to identify other Telegram discussions that could also be scrapped for discussions relevant to the invasion of Ukraine. Thus, the overall strategy for collecting data in an information environment where social media sites have walled gardens that prohibit searching across the entire platform and where a profusion of websites, especially news sites, exists is to leverage crowdsourcing combined with snowball sampling to expand the network of links, and thus more fully collect relevant, available data. This methodology addresses the main problem in these environments: Finding links to the data itself. Furthermore, since the collection was done in a more focused manner than trying to scrape everything and was collected in a way that maximizes relevance to social media users, it will also better identify more relevant information for the researcher.

Limitations and Future Work

It is important to note that there are some challenges associated with this strategy. Firstly, once the links to the data have been found, the data still needs to be scraped. This can pose a particular challenge for sources such as news websites, which often lack APIs, requiring HTTP requests and possibly rendering elements like JavaScript. Tools such as requests in Python (Chandra and Varanasi 2015) and using a headless browser such as Selenium to access dynamic sites (Gheorghe, Mihai, and Dârdală 2018) can aid in getting access to the data on the webpage. This data still has to be parsed (Uzun, Yerlikaya, and Kirat 2018) and often requires additional text-level processing to make it usable. So, while there are tools available to scrape and process data, given a URL, these tools still have to be configured to handle different types of websites and frequently result in bespoke data collection, rather than the more generically-formatted and applicable data that would come from an API.

Furthermore, as more mainstream social media sites tighten their access, finding seed links becomes increasingly difficult, as mainstream social media sites are typically reliable sources of seed links to walled gardens or specific news sites. While these challenges can be overcome with tools like headless browsers for scraping, they do present distinct technical hurdles that must be addressed and can potentially slow down the data acquisition process.

Another limitation in this method arises from the nature of snowball sampling: it is possible that the chosen starting seeds could bias the analysis to only a subset of the total conversation on a given topic (Kirchherr and Charles 2018). This is especially problematic on social media platforms and online communities where linking to other viewpoints is less common (e.g., Reddit). To mitigate this issue, it is important to find a sample of seeds that represent the spectrum of viewpoints on a topic. Fortunately, this issue is less of a problem on platforms like Telegram, where previous work has found robust linking between channels with contradictory viewpoints (Kloo and Carley 2023).

Finally, it is important to note that assessing the coverage, reliability, and representativeness of collecting data using the proposed methodology remains a challenge. while thee has been much work on assessing the quality of samples produced by snowball sampling, or link tracing, over the years, most of that work is attempting to address representativeness in the face of data volume challenges (Kirchherr and Charles 2018; Lecy and Beatty 2012; Heckathorn and Cameron 2017; Chan 2020). More specifically, previous work on assessing snowball sampling have sought to understand how to create a representative sample when there is too much data to collect; how little data does one need to collect to still have a representative smaple. In our work with this methodology, we have found that while the volume of data (i.e., URLs to scrape) can be a challenge, there is also a challenge of salience of the data; not all of the links being shared in online conversations are salient to the topics of discussions taking place. Thus, the data collected by this method also frequently need to be assessed for their salience, often through something like topic modeling and filtering. Finding better ways to assess data from the proposed method-

ology remains a challenge and a basis for future work and discussion.

References

Axel Bruns; and Bruns, A. 2019. After the 'APIcalypse': social media platforms and their fight against critical scholarly research. *Information, Communication & Society*, 22(11): 1544–1566. MAG ID: 2956617434.

Balakrishnan, V.; Ng, W. Z.; Soo, M. C.; Han, G. J.; and Lee, C. J. 2022. Infodemic and fake news–A comprehensive overview of its global magnitude during the COVID-19 pandemic in 2021: A scoping review. *International Journal of Disaster Risk Reduction*, 78: 103144.

Chan, J. T. 2020. Snowball sampling and sample selection in a social network. In *The Econometrics of Networks*, 61–80. Emerald Publishing Limited.

Chandra, R. V.; and Varanasi, B. S. 2015. *Python requests essentials*. Packt Publishing Birmingham, UK.

Cruickshank, I.; Ginossar, T.; Sulskis, J.; Zheleva, E.; and Berger-Wolf, T. 2021. Content and dynamics of websites shared over vaccine-related tweets in COVID-19 conversations: Computational analysis. *Journal of Medical Internet Research*, 23(12): e29127.

Cruickshank, I. J.; and Carley, K. M. 2020. Clustering analysis of website usage on twitter during the covid-19 pandemic. In *Annual International Conference on Information Management and Big Data*, 384–399. Springer.

Durbin, B. 2023. X changes its API to retire legacy tiers and endpoints.

Gao, H.; Barbier, G.; and Goolsby, R. 2011. Harnessing the crowdsourcing power of social media for disaster relief. *IEEE intelligent systems*, 26(3): 10–14.

Gheorghe, M.; Mihai, F.-C.; and Dârdală, M. 2018. Modern techniques of web scraping for data scientists. *International Journal of User-System Interaction*, 11(1): 63–75.

Ginossar, T.; Cruickshank, I. J.; Zheleva, E.; Sulskis, J.; and Berger-Wolf, T. 2022. Cross-platform spread: vaccine-related content, sources, and conspiracy theories in YouTube videos shared in early Twitter COVID-19 conversations. *Human vaccines & immunotherapeutics*, 18(1): 1–13.

Heckathorn, D. D.; and Cameron, C. J. 2017. Network sampling: From snowball and multiplicity to respondent-driven sampling. *Annual review of sociology*, 43: 101–119.

Kirchherr, J.; and Charles, K. 2018. Enhancing the sample diversity of snowball samples: Recommendations from a research project on anti-dam movements in Southeast Asia. *PloS one*, 13(8): e0201710.

Kloo, I.; and Carley, K. M. 2023. Social cybersecurity analysis of the telegram information environment during the 2022 invasion of Ukraine. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*, 23–32. Springer.

Lecy, J. D.; and Beatty, K. E. 2012. Representative literature reviews using constrained snowball sampling and citation network analysis. *Available at SSRN 1992601*.

Leighton, K.; Kardong-Edgren, S.; Schneidereith, T.; and Foisy-Doll, C. 2021. Using social media and snowball sampling as an alternative recruitment strategy for research. *Clinical simulation in nursing*, 55: 37–42.

Lepird, C. S.; and Carley, K. M. 2023a. Automated Pink Slime Detection from Social Network Features. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*.

Lepird, C. S.; and Carley, K. M. 2023b. Comparison of Online Maneuvers by Authentic and Inauthentic Local News Organizations. *arXiv preprint arXiv:2312.07613*.

Ng, L. H.; and Taeihagh, A. 2021. How does fake news spread? Understanding pathways of disinformation spread through APIs. *Policy & Internet*, 13(4): 560–585.

Ng, L. H. X.; Cruickshank, I. J.; and Carley, K. M. 2022. Cross-platform information spread during the January 6th capitol riots. *Social Network Analysis and Mining*, 12(1): 133.

Uzun, E.; Yerlikaya, T.; and Kirat, O. 2018. Comparison of python libraries used for web data extraction. *Journal of the Technical University-Sofia Plovdiv Branch, Bulgaria*, 24: 87–92.