**Introduction:** Researchers have recently uncovered a dark underbelly of the news and information ecosystem, where users spread disinformation, or false information, on social media [1]. In today's world, where 67% of Americans report they receive some fraction of their news from social media [2], disinformation has the dangerous potential to influence and persuade the mind of the public. Further, these disinformation campaigns have been hypothesized to leverage *automation* to artificially amplify their message online [1, 3], making it appear that many distinct users share their viewpoints. The security community has studied the impact of automation in other flavors of Internet abuse, such as spam [4] and phishing [5]; my work aims to leverage our understanding of Internet abuse to characterize automation in the information ecosystem.

**Project I: Characterizing Automated Campaigns through Network Infrastructure.** Though prior work has studied automation in the information ecosystem, it has largely done so by either analyzing the content of posts or by investigating the properties of the underlying social network graph [1, 3]. To the best of my knowledge, no research has investigated automation in the information ecosystem through its underlying network infrastructure, such as the source IP or source network of automated messages. Studying network infrastructure will provide lower-level insight into the technical operation of disinformation campaigns, which may be more valuable in mitigating their harmful effects than content based analysis. Thus, I propose leveraging network infrastructure in characterizing automated information campaigns on social media. For this project, I will primarily focus on Twitter, due to the public-by-default nature of tweets.

First, I plan to partner with Twitter to continuously retrieve a representative sample of all public tweets. Then, I will identify the IP addresses used to publish tweets of interest. Prior work has hypothesized automation by aggregating tweet metadata; I will validate this hypothesis by additionally aggregating on network infrastructure. For example, if identical tweets are published at the same time, by different users, *and* from the same network infrastructure, this serves as a reasonable indicator of tweet automation.

After identifying automation, I will characterize the networks that the tweets are sent from. I will identify the distribution of IPs, and subsequent networks that contribute the most to automated campaigns. Further, I will identify the classes of networks utilized, for example, a commercial ISP versus a government-operated network. This provides insight into which networks are responsible for the most automation, which is currently opaque to researchers in this field.

Finally, I will tie our analysis of automation back to disinformation campaigns. One example of these campaigns is currently targeting the White Helmet movement in Syria [6], which I can investigate as a case study. I will identify automatically generated tweets that mention the White Helmets, and further investigate the users that generate automated content. Finally, I will study the automated content itself, to characterize what narratives are being amplified on this topic.

**Project II: Measuring the Effectiveness of Automated Campaigns.** Though research have shown that automated campaigns exist in the ecosystem, measuring the effectiveness of these campaigns remains a challenge. Prior work has investigated dispersion of automated content, either through relationships in social graphs or their appearance across disparate social media platforms [3]. Unfortunately, this does not provide any way of identifying how many people automated content reaches. To address this challenge, I propose drawing on three prior Internet abuse identification techniques.

First, I will leverage Domain Name System (DNS) data to determine the lookup volume of amplified domains. I can observe change in lookup volumes longitudinally, and ultimately correlate

significant changes over time to automated campaigns identified from Project I. I will further use these changes as a proxy to quantify the change in impressions at the time an automated campaign is launched.

Second, I will work with industry partners to study their network logs. From the logs, we can identify the number of unique IP addresses that interact with amplified content, which lends to understanding *where* automated campaigns have the most effect. For example, we can identify the networks that consume the most amplified content, and further, determine the locations where automated campaigns have the largest impact.

Finally, I will carefully design ethical experiments that deploy artificial, automated disinformation campaigns in the wild. I will use this to determine the factors that have an impact on the number of impressions garnered by such campaigns, such as the such as the size of the underlying botnet or the category of amplified content (news, politics, etc.). Ultimately, this will provide new insight into how new automated campaigns may be run, and can help inform preventive defenses against them.

**Project III: Building Tools for End-Users.** Our analysis of automated campaigns in the information ecosystem will result in new insight into their technical operations. However, such insight is lost without reporting the results back to users. To this end, I will build end-user tools and systems that can aid in the general identification of automated information on the Internet.

Building off prior community experiences deploying Bobble [7], a browser extension to track personalized content, I plan to implement a similar tool to reach end users. I will first build and deploy a new browser extension that will flag social media content that we detect was created automatically. In addition, this extension will enable users to report new instances of content they feel may be artificially amplified, and will serve as a secondary dataset for tuning our classification models. The crowdsourced data will further provide insight into the dispersion and impressions of automated content from the perspective of the client, which we would previously have little-to-no visibility into.

**Intellectual Merit:** My research will provide deeper insight into the role of automation in launching mass disinformation campaigns. In characterizing automation, my work will lend to other security challenges, such as tracking information provenance on the web and identifying maliciously motivated networks. Finally, my research will provide a look into how disinformation campaigns are carried out today, which will guide future work carried out by psychologists, political scientists, and other interdisciplinary researchers in this space.

**Broader Impact:** I aim to partner closely with abuse teams in industry to implement my research, thereby closing the gap between academic knowledge and real-world usage. Based on my research, I will further build new, usable tools to educate users in identifying disinformation in their regular browsing experience. Finally, I will open-source my research tools to the community for future researchers to build new systems off of. Ultimately, in characterizing automation in the information ecosystem, I seek to mitigate the harm disinformation has on our autonomy, society, and our democracy.

**References** [1] K. Starbird. Information wars: A window into the alternative media ecosystem. [2] E. Shearer and J. Gottfried. News use across social media platforms 2017. [3] S. Zannettou, T. Caulfield, E. D. Cristofaro, N. Kourtellis, I. Leontiadis, M. Sirivianos, G. Stringhini, and J. Blackburn. The web centipede: Understanding how web communities influence each other through the lens of mainstream and alternative news sources. [4] C. Kanich, C. Kreibech, K. Levchenko, B. Enright, G. M. Voelker, V. Paxson, and S. Savage. Spamalytics: An empirical analysis of spam marketing conversion. [5] M. Cova, C. Kruegel, and G. Vigna. There is no free phish: An analysis of free and live phishing kits. [6] E. G. Ellis. Inside the conspiracy theory that turned syria's first responders into terrorists. [7] X. Wang, W. Meng, D. Doozan, N. Feamster, W. Lee, and A. C. Snoeren. Exposing inconsistent web search results with bobble.