Following Their Footsteps: Characterizing Account Automation Abuse and Defenses

Louis F. DeKoven
University of California, San Diego
ldekoven@cs.ucsd.edu

Trevor Pottinger
Facebook
tpott@fb.com

Stefan Savage
University of California, San Diego
savage@cs.ucsd.edu

Geoffrey M. Voelker
University of California, San Diego
voelker@cs.ucsd.edu

Nektarios Leontiadis
Facebook
leontiadis@fb.com

ABSTRACT

Online social networks routinely attract abuse from for-profit services that offer to artificially manipulate a user's social standing. In this paper, we examine five such services in depth, each advertising the ability to inflate their customer's standing on the Instagram social network. We identify the techniques used by these services to drive social actions, and how they are structured to evade straightforward detection. We characterize the dynamics of their customer base over several months and show that they are able to attract a large clientele and generate over $1M in monthly revenue. Finally, we construct controlled experiments to disrupt these services and analyze how different approaches to intervention (i.e., transparent interventions such as blocking abusive services vs. more opaque approaches such as deferred removal of artificial actions) can drive different reactions and thus provide distinct trade-offs for defenders.

CCS CONCEPTS

• **Security and privacy** → *Social network security and privacy*;

ACM Reference Format:
Louis F. DeKoven, Trevor Pottinger, Stefan Savage, Geoffrey M. Voelker, and Nektarios Leontiadis. 2018. Following Their Footsteps:, Characterizing Account Automation Abuse and Defenses. In *2018 Internet Measurement Conference (IMC '18), October 31-November 2, 2018, Boston, MA, USA.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3278532.3278537

1 INTRODUCTION

Social media, as with all forms of mass communication, provides a platform whereby a single message can reach large audiences. However, the reach of any given message is determined by the popularity of the user who publishes it. Concretely, users with more followers are able to reach larger audiences with their posts and thus can be seen as carrying more "weight" in some abstract social hierarchy. Since this standing is directly monetizable via advertising, it is unsurprising that this aspect of social media has attracted organized abuse. Indeed, the medium has engendered a large underground service market that focuses on bypassing the organic nature of social relationships and instead advertises the ability to create artificially enhanced social network status in exchange for payment.

In this paper we explore this phenomena in the context of the popular Instagram photo-sharing service. To wit, searching for "Instagram likes" in a search engine will produce pages of sites with inducements such as "Buy Instagram Likes from $2.97 only!" or "Instant Instagram Likes — 100% Real & Genuine Likes". However, the precise mechanism by which such services ply their trade is unclear and, in fact, simplistic "bot-based" approaches (whereby a service creates fake accounts and uses them to initiate social actions to customer content) are easy to detect and filter. In our work, we focus on the more sophisticated segment of this market, Account Automation Services (AASs) in which users provide their Instagram credentials to third party actors who, in turn, use those credentials to perform actions on the user's behalf in a manner that violates Instagram's Terms of Use [13].

We have explored these services through a variety of techniques. Using a broad array of independent "honeypot accounts" we engaged (on behalf of these accounts) with five large account automation services: Instalex, Instazood, Followersgratis, Boostgram and Hublaagram. By requesting a range of "social actions" from each AAS, and then monitoring activity to and from the associated accounts, we inferred the mechanisms each service uses to achieve its ends. Notably, we distinguish two distinct techniques — collusion networks and reciprocity abuse — used to artificially create social connectivity. Using our service characterizations we were then able to identify all accounts used by customers of each service. Collecting data on this corpus over several months, we were able to characterize the dynamics of their customer populations and the underlying revenue of each business. Finally, we performed controlled experiments to evaluate different kinds of interventions (*e.g.*, blocking such services from accessing Instagram vs. removing their actions at a future date) and the reactions each kind of intervention evoked from the services and their customers.

We believe our work is the most comprehensive study of this kind to date on Instagram, and that our analysis provides several insights that were not previously understood or lacked empirical validation in the broader space of social network abuse:

- **Social action laundering.** We identify two techniques designed to artificially create social actions while evading traditional detection mechanisms. The first, *reciprocity abuse*, leverages the tendency of some users to issue complementary `follows` or `likes` in response to an unknown user

IMC '18, October 31-November 2, 2018, Boston, MA, USA

ACM ISBN 978-1-4503-5619-0/18/10.
https://doi.org/10.1145/3278532.3278537

`following` them or `liking` their content. This reciprocity effect allows services to quickly inflate the `follower` or `like` counts of their customers by automating *outbound* actions to a curated set of recipients.

The second approach, *collusion networks*, uses the entirety of a service's population to orchestrate the exchange of social actions. Thus, each customer account is used to issue `follows` or `likes` to other customers, and they in turn receive inbound actions from yet other customers (similar, in principal, to the notion of a mix network [4]).

- **Commercial scale.** We find that these services are quite successful as business entities and we estimate the gross revenue among three of the five AASs alone to be over $1M *per month*. Moreover, we show that long-term customers (*i.e.*, customers who repeatedly contract for services over multiple months) provide the lion's share of these proceeds (*i.e.*, that the core set of customers is stable and customer churn is modest).
- **Intervention impacts.** We experimentally demonstrate that transparent interventions (*e.g.*, blocking actions from a given account automation service) quickly provokes adversarial adaptation, while deferred interventions (*e.g.*, removing service actions a day later) is far more likely to go unanswered. Somewhat unintuitively, our results suggest that related abuse interventions will be most effective and long-lived precisely when they do not visibly undermine the business model of the abusive service.

In the remainder of this paper we provide background on how such social networks operate and are abused, describe the set of AASs we explored, and provide a detailed description of our measurement methodology. We provide an analysis of both user dynamics and service revenue, and then describe a series of controlled intervention experiments that explore how for-profit service abuse businesses respond to different varieties of disruption.

2 BACKGROUND

Online social networks (*e.g.*, Twitter, Facebook, Instagram, Youtube, Snapchat, etc.) are targeted by abusers that engage in activities spanning from selling fake actions to hijacking user accounts. We are not the first to identify this phenomenon, and other researchers have characterized a range of such practices that we build on in our own work. Javed *et al.* characterized generic traffic exchange services that provide customers with inflated view counts—including for social media—from large pools of IP addresses, and find many exchanges which pay users in return for views to their content [18]. This work establishes both the commercial nature of such abuse and the use of live humans as traffic sources. Hooi *et al.* develop a bipartite graph algorithm to detect abusive actions on the Twitter follower-followee graph, where miscreants may camouflage their abusive actions by producing actions to non-customers [10]. Again, this work identifies the use of organic (i.e., non-bot) accounts as a critical challenge in social network abuse and uses statistical techniques to try to distinguish legitimate and illigitmate actions from such accounts.

Other researchers have tried to overcome this issue by using honeypot accounts to crisply identify abuse targeting across a range of social networks including MySpace, Twitter, and Facebook [1, 19, 26, 29, 30]. Moreover, in some cases, this data has then been used to successfully train classifiers to identify those accounts complicit in collusion networks [1, 29]. Our work builds on both of these techniques—the use of honeypots to obtain abuse data and using this data to train abuse classifiers—in our analysis of Instagram abuse.

The honeypot approach has also been combined with active purchasing from third-party services to investigate commercial abuse. For example, De Cristofaro et al's analysis of Facebook services [5] and Stringhini et al's analysis of Twitter following services [27] both use this approach and characterize the nature of the fraudulent social networks they find. Our work is distinct, not only due to the different social network examined (Instagram), but also because we focus on more complex (i.e., non-bot) forms of abuse in our work. As well, we are able to provide a grounded analysis about service revenue that informs how we consider the nature of the threat and focused experiments exploring the impact of different interventions.

Mislove *et al.* identified the existence of high degrees of reciprocated actions within online social networks (*e.g.*, Flicker, YoutTube, etc.) which, a decade later, forms the basis for the reciprocity abuse we identify in this work [22]. Finally, most closely related to our work is that of Farooqi *et al.* who describe a collusion network abusing third-party application OAuth tokens on Facebook, and the results of large-scale network-level blocking of the organizations behind this activity [7]. Our work brings a related analysis to a distinct social network and extends it by analyzing reciprocity abuse as well as collusions networks, quantifying the underlying business and revenue model for multiple abuse groups, and performing active experiments with finer-grained (i.e., account-level) interventions.

For this paper, we focus squarely on Instagram, a popular online social network structured around sharing and discussing photos posted by its 800 million users [12]. In normal use, each Instagram user can upload photos and videos, apply visual filters and tag photos with hashtags. A user's followers will see the media the user has posted, and can interact by `liking` the media and posting comments. Thus, users with more followers will have their content exposed to a broader audience and will receive on average more interactions.

Typically, differences in social status (*e.g.*, the number of `likes` per photo, `followers`, etc.) are an organic byproduct of each user's own authentic activity. However, in addition to the implicit psychological factors that drive users to desire increased social standing, there can be strong economic incentives as well. Notably, after reaching a social status commonly referred to as an "influencer", outside businesses may offer to pay users thousands of dollars in exchange for posts (*e.g.*, for marketing purposes) [21, 24]. It is a popular belief in this community that, to become an influencer, a user of Instagram needs an account with both a high engagement (*i.e.*, a large number of other Instagram users that interact with posted content), and thousands of followers [21]. The potential for such inducements leads some users to pursue increased social status via abusive means, and gives rise to third-party services that perform this function for a fee. Indeed, such services formalize this notion and promote a metric called the "engagement rate" to evaluate the quality (and hence potential profitability) of an influencer [16]. They argue that users should try to maximize this metric:

$$ER = \frac{\text{Number of likes \& comments}}{\text{Number of followers}}$$

and commonly offer to manipulate one or more of its components as a key aspect of their service offering (with one such service claiming that each $1 spent produces a return of $6 in marketing revenue).

One approach for achieving this end is to create a range of synthetic Instagram accounts and use them to `follow` the accounts of paying customers, `like` their content, and so on. However, this kind of purely synthetic account manipulation can be easy to detect. Indeed, over the last year Instagram has worked to disrupt a range of popular bot services including Instagress, MassPlanner, PeerBoost, InstaPlus, and FanHarvest [9, 20, 23, 28]. The more sophisticated players in this ecosystem perform "account automation" whereby their customers provide access to their Instagram login credentials, and the service performs actions on their behalf. In fact, Instagram provides a public OAuth-based API that allows a Web site to perform actions on behalf of users that grant permission. However, this API is rate limited in a manner that precludes broad abusive use. Thus, most commercial account automation services bypass these limitations by reverse engineering the private API used by the Instagram mobile client and generating spoofed requests to appear as valid mobile client actions.

3 ACCOUNT AUTOMATION SERVICES

Based on our observations, AASs use two distinct approaches to achieve their ends: (i) Reciprocity Abuse and (ii) Collusion Networks. The former aims to provide *authentic* actions (*i.e.*, `likes`, `follows`, etc.) to their customer's Instagram account, while the latter provides customers with *inauthentic* actions to their Instagram account. In this section we describe each approach, and then detail the particular set of services we studied in this effort.

3.1 Reciprocity Abuse

Reciprocity Abuse AASs provide their customers with organic actions from other Instagram user accounts by exploiting the concept of social *reciprocity*. For example, when Instagram user $A1$ receives an (inbound) action from Instagram user $B2$, $A1$ will be notified in real-time about $B2$'s action, and $A1$ may reciprocate by performing an action to user $B2$. This "you follow me, I follow you" behavior is an organic response taken by some subset of Instagram users. Reciprocity Abuse AASs abuse this behavior by automating large numbers of (outbound) actions from their customer's Instagram account in the hope that a subset of users receiving an action will return the favor in kind — thus providing their customer with inbound actions, such as `follows`.

3.2 Collusion Networks

By contrast, Collusion Network AASs provide their customers with inbound *inauthentic* actions on their Instagram accounts. A collusion network is a group of Instagram accounts used in concert to orchestrate actions to one another. Accounts participating in the collusion network will produce outbound actions to other accounts in the network, as well as receive inbound actions from the network. Customers of Collusion Network AASs are hoping to strictly increase the number of actions on their Instagram account and they are willing to have their account used in the same manner on behalf of others to serve this goal.

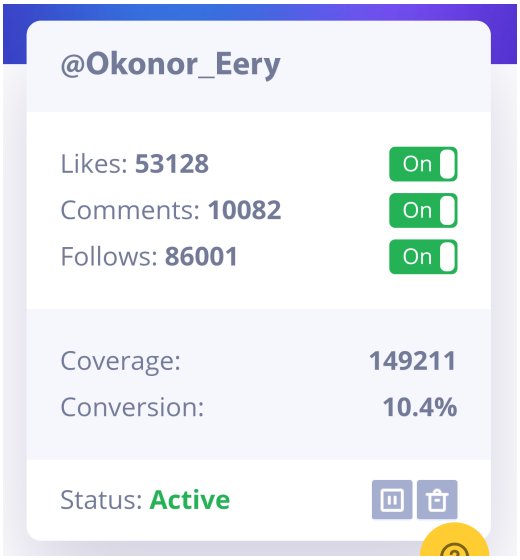

Figure 1: Instalex Web site providing an example account control panel with action counts performed on Instagram.

3.3 Studied services

We study five popular AASs in detail that we discovered through a combination of searching popular underground forums for popular recommendations from the community, together with repeated complaints from Instagram users caused by unsolicited AAS advertisements. Three use the reciprocity abuse approach (Instalex, Instazood and Boostgram), while the other two implement collusion networks (Hublaagram and Followersgratis). For each service, we explored its Web site in fall 2017 to understand the registration process, what features are offered, and the advertised business model [3, 8, 11, 14, 17]. Figure 1, for example, shows a screenshot of the Instalex customer control panel. During this process, we also discovered that the Instalex and Instazood services were independently operated franchisees of the same parent organization (which offers franchising services ranging from $1,990 to $30,990 per month [15]). Since they appear to be operated independently, we evaluate these two services separately until Section 5 where we combine the two services when we cannot separate their actions.

3.3.1 Registration Process. Both Reciprocity Abuse AASs and Collusion Network AASs produce automated activity from the Instagram accounts of their customers. Therefore, a required step when registering for any AAS is for the customer to provide their Instagram account credentials (*e.g.*, in contrast to abuse methods where the AASs depends on the ability to use customer OAuth tokens [7]). By sharing their Instagram credentials the customer gives an AAS *full* control over their Instagram account, while resetting the password revokes AAS access to the account.

Table 1 shows the different AASs by name, service type, and what services are available to customers. All offer `like` and `follow` services, 60% offer `comment` and `unfollow` services, and 40% offer `post` services. It comes as no surprise that every AAS offers at a minimum `likes` and `follows` as these are the most frequent actions

Service	Like	Follow	Comment	Post	Unfollow
Reciprocity Abuse AASs					
Instalex	★	★	★		★
Instazood	★	★	★	★	★
Boostgram	★	★		★	★
Collusion Network AASs					
Service	Like	Follow	Comment	Post	Unfollow
Hublaagram	★	★	★		
Followersgratis	★	★			

Table 1: Services offered to customers of Reciprocity Abuse AASs and Collusion Network AASs.

on Instagram. Some AASs provide `comment` and `post` services as additional ways for their customers to attract other Instagram users to engage with their content. Lastly, all Reciprocity Abuse AASs provide `unfollow` services that allow their customers to remove the outbound `follows` performed by the AAS in an effort to retain only the inbound `follows` they receive.

Many Reciprocity Abuse AASs allow their customers to target groups of Instagram accounts that will receive automated actions, allowing their customers to obtain reciprocated actions from users with common interests. Customers can provide either a list of Instagram users, or a list of hashtags to narrow the accounts that a AAS will interact with. When signed into a Collusion Network AAS, customers are typically given the option to request a specific *type* and *quantity* of inbound actions (*e.g.*, 2,000 `likes`, etc.) to other customers of the network, but cannot specify the interests of accounts they receive actions from.

3.3.2 AAS Business Model. The primary revenue source across the studied AASs is customer payments for the services they offer.[1] In turn, there are two different techniques used by AASs to attract customers in the hope that they become paying customers: trial periods, and free services.

First-time customers of Reciprocity Abuse AASs are commonly offered a free variable-length trial period. During the trial period customers have access to all of the service's features. However, as soon as the trial period expires the service is discontinued, and if the customer wants to continue service they are required to pay. Reciprocity Abuse AASs have a relatively straightforward cost structure where customers pay for each of their Instagram accounts to gain full use of the service for a specified time period. Table 2 presents the free and paid service options for customers of the Reciprocity Abuse AASs we study.

Collusion Network AASs offer customers the ability to periodically request small quantities of actions onto their Instagram account for "free". Soon after a customer provides their Instagram credentials the service will begin to use the account in the collusion network. Hublaagram provides free `likes`, `follows`, and `comments`, while Followersgratis only offers free `follows`. Free service, though, is rate-limited; Hublaagram, for instance, has a 30-minute timeout between requests. Naturally, both Collusion Network AASs encourage customers to pay money to receive a larger quantity of actions. We present the different paid service options for Hublaagram and Followersgratis in Tables 3 and 4, respectively.

[1] There is also a minor revenue stream arising from advertisements shown to customers, but it does not appear to be significant by comparison (Section 5.2).

Service	Trial Days	Min Paid Days	Cost
Instalex	7 days	7	$3.15
Instazood	3 days	1	$0.34
Boostgram	3 days	30	$99

Table 2: For Reciprocity Abuse AAS we show the free trial length, the minimum number of days that service can be purchased for, and the corresponding cost per Instagram account.

Description	Cost	Duration
No collusion network	$15	Life
2,000 Likes	$10	Immediate
5,000 Likes	$20	Immediate
10,000 likes	$25	Immediate
250 − 500 Likes	$20	Month
500 − 1,000 Likes	$30	Month
1,000 − 2,000 Likes	$40	Month
2,000 − 4,000 Likes	$70	Month

Table 3: All per-account costs for Hublaagram services. Hublaagram allows customers to pay a one-time fee that prevents their Instagram account from participating in the collusion network. Services with an immediate duration are applied as fast as possible to a single post, and services with a month duration have the purchased quantity of `likes` applied to *each* new photo posted on the account throughout the month.

Description	Cost	Duration
500 Follows (300 free likes)	$3.15	1 Day
1,000 Follows (500 free likes)	$5.25	1 Day
500 Likes (250 free likes)	$2.10	Instant
500 Likes (500 free likes)	$5.25	Fast

Table 4: The Followersgratis payment options. With `likes`, customers who select the less expensive option receive `likes` from Instagram accounts located around the world on five different photos. The more expensive `like` option provides `likes` from Instagram accounts located in Indonesia, and the `likes` are spread across ten photos. The duration for `likes` is specified explicitly on the Followersgratis Web site without explanation.

4 USER EXPERIENCE

In this section we evaluate the experience of using Account Automation Services from a user's perspective using a collection of fully-instrumented honeypot accounts.

4.1 Methodology

To identify abusive actions generated by the AASs, we registered multiple distinct honeypot accounts with each service described in Section 3.3. Thus, for each account, we register it with an AAS, request that the service perform either inbound or outbound actions on the account, and then monitor the resulting actions. Since they neither generate nor receive organic actions, honeypot accounts are particularly useful because we can attribute all activity to the linked AAS. We describe our methodology for using honeypot accounts in more detail below.

4.1.1 Account Types. We developed a honeypot account framework to programmatically manage a large number of Instagram accounts. Our framework supports campaign-specific accounts, account creation, posting content, deletion, and data collection of all inbound and outbound actions on the account. When deleting a honeypot account, all actions to or from the account are eventually removed from Instagram.

For each service, we created two different types of honeypot accounts to determine if AASs differentiate between fake or real-looking Instagram accounts (they do not), and if there is a difference between reciprocated action rates from Instagram users that receive an outbound action from AASs (there is; more below in Section 4.3).

The two types of honeypot accounts we register are "empty" and "lived-in" accounts. Empty accounts contain the minimum information required to use all of the AASs that we study. In particular, we populate honeypot accounts with 10 or more photos from one of the following categories: dogs, cats, lizards, and food. Lived-in accounts, in addition to having uploaded photos, are fully populated Instagram accounts with a profile picture, biography, and name, all unique. Lived-in accounts follow $10 - 20$ high-profile Instagram accounts (>1M followers), but do not themselves have followers when created. Beyond enrolling them in the AAS services, we do not use them to perform actions on Instagram after being created.

4.1.2 Account Registration. We registered 10 honeypot accounts for every service type offered by each AASs listed in Table 1, specifying that the account be used *only* for that service type. For example, as Instalex offers three different services, we registered 30 accounts in the service. Among each set of 10 accounts, nine are empty and one is lived-in. In total we registered over 150 honeypot accounts during the course of a month of manual registration effort. Moreover, some of our accounts engaged with the free services offered by each AAS while others explicitly paid for contracted services. For AASs that require target information for particular actions (*e.g.*, targets of `likes` and `follows`), we created a static list of hashtags and Instagram accounts that could be used in common. We chose relatively high-profile hashtags and Instagram accounts (*e.g.*, having more than 1M followers) to reduce the impact of the temporary actions produced from our honeypot accounts. We also made a point to use a diverse set of commercial and residential IP addresses when accessing each AAS's site in the unlikely event that any of the services actively monitor and correlate connections to their site. Finally, we deleted our honeypot accounts after the measurement period, which removed all of their actions from Instagram.

4.1.3 Attribution. When using honeypot accounts with AASs, we attribute the activity on those accounts solely to their involvement in the AASs. To rule out the possibility that the activity could be due to other users of Instagram, we used a separate set of 50 inactive honeypot accounts to establish a baseline of background activity on Instagram. The inactive accounts are not registered with an AAS, and we never used them to produce actions that are visible to other users of Instagram.

For each account we similarly uploaded at least 10 photos at the time of creation. We then actively monitored whether any inbound action (*i.e.*, `likes`, `follows`, etc.) took place on these accounts. For the duration of our study, we did not observe any activity on any of the inactive honeypot accounts. As a result, for the honeypot accounts we register with AASs we attribute all activity on those accounts to their involvement with the services.

4.2 How Accounts Are Used

Using the honeypot accounts, we examine how AASs use the accounts registered with their services.

Since customers provide their Instagram credentials to an AAS during registration (Section 3.3.1), it is possible for the AAS to abuse the Instagram account to produce additional, potentially undesired actions. We compared the types of actions we requested with the types of actions the services actually performed with our accounts (*e.g.*, when requesting `likes` does a service use the account for anything other than `like` actions?). The services all perform as advertised. Across the AASs we study, they only perform actions of the type we requested, and no AASs used our accounts to produce visible un-requested actions.

In later analyses in Section 5, such as estimating revenue, it is important to distinguish between users using the free trial periods on services and those users paying money for service. Although the services advertise the lengths of their trial periods (Table 2), we also experimentally evaluated their durations using the honeypot accounts. Trial service starts immediately, with our accounts becoming active within minutes of requesting free service. And with one exception, we confirmed that free trial service lasts for the advertised period, and that activity with accounts stops no more than 12 hours beyond the expected end time. Instazood, however, advertises a three-day trial period, yet all of our honeypot accounts received seven days of trial service. As a result, for Instazood we assume that trial period activity is seven days.

4.3 Quantifying Reciprocation

As a final experiment we use our honeypot accounts to measure the probability that an outbound `like` or `follow` will spontaneously generate a reciprocated action. Previous work has shown how collusion networks use their control over the accounts in the network to serve as both the source and target of actions [7]. In contrast, Reciprocity Abuse AASs fundamentally rely upon natural social behavior in online networks to fulfill their customer requests. As discussed in Section 3.1, these services produce outbound actions from user accounts under their control, but the targets of these actions are other Instagram accounts that are *not* under the control of the service. The underlying assumption is that, for each action, there is some probability that the target of the action will naturally reciprocate with a similar action. With a sufficiently high volume of outbound

Service	Outbound	Inbound	
		Likes	Follows
Boostgram (E)	Likes	1.5%	0.1%
Instalex (E)	Likes	2.1%	1.4%
Instazood (E)	Likes	2.1%	0.2%
Boostgram (L)	Likes	3.9%	0.2%
Instalex (L)	Likes	3.7%	1.8%
Instazood (L)	Likes	3.5%	0.4%
Boostgram (E)	Follows	0.0%	10.3%
Instalex (E)	Follows	0.0%	12.8%
Instazood (E)	Follows	0.0%	13.0%
Boostgram (L)	Follows	0.0%	12.0%
Instalex (L)	Follows	0.0%	13.7%
Instazood (L)	Follows	0.0%	16.1%

Table 5: The probability of receiving a reciprocated inbound action given an outbound action of a specific type. For each service, we show the reciprocation ratio for both empty (E) and lived-in (L) honeypot accounts.

actions, these services can then organically induce reciprocating actions to satisfy their customer requests.

Table 5 shows the probability of receiving a reciprocated action given an outbound `like` or `follow` for the three Reciprocity Abuse AASs. We separate the results for the two different kinds of honeypot accounts, empty (E) and lived-in (L). For example, generating an outbound `like` with our empty Boostgram honeypot accounts has a 1.5% chance of inducing a reciprocating `like` and a 0.1% chance of inducing a reciprocating `follow`. These results quantify the reciprocity effect of users on Instagram, and from them we make a number of observations.

First, the reciprocation rates are for the most part very consistent across the services. Although Instalex and Instazood are franchises of the same service, they also exhibit reciprocation rates that are similar with those on Boostgram. These results are consistent with these services tapping into fundamental underlying online social behavior on Instagram. Moreover, the reciprocation rates are relatively high for `follows`. For just 6–10 outbound `follow` actions, our honeypot accounts receive a new inbound `follow` from a real user. (In Section 5.3, we show that the services appear to specifically target users who are more likely to respond to inbound `follows` to increase the probability of reciprocation.)

The one anomaly is inbound `follows` to outbound `likes` on Instalex, which has a reciprocation rate many times greater than the other services. Exploring further, though, we found no significant features in the accounts targeted by Instalex compared to the other services that might explain the difference: The inbound actions come from hundreds of autonomous systems, the time between when the actions take place and when the honeypot account was registered in the service is uniformly distributed throughout the trial period, the inbound actions come from dozens of countries, etc. As a result, we currently do not have an explanation for this one difference.

Second, users primarily reciprocate with the same action, e.g., Instagram users reciprocate with a `like` when receiving a `like` from one of our accounts. Much less often, users will reciprocate to an outgoing `like` by `following` one of our accounts (an order of magnitude less often for Boostgram and Instazood). And users never reciprocate with `likes` when `followed` by one of our accounts.

Finally, Instagram users are sensitive to the differences in honeypot accounts. Confirming expectations, empty accounts have a significantly smaller probability of receiving reciprocal inbound actions than lived-in accounts, particularly for `likes`. Lived-in accounts range from 1.6× as likely on Instazood to 2.6× as likely on Boostgram to generate inbound `likes`. This difference confirms the utility of more realistic honeypot accounts.

5 BUSINESS PERSPECTIVE

Our honeypot accounts gave us insight into the AASs from a user's perspective. They were also valuable in providing us with ground-truth on AAS activity, which we were then able to use to identify all activity generated by all Instagram accounts used by the AASs. Based on features gathered from our honeypot accounts, such as the type of action (*e.g.*, `like`, `follow`, account login, etc.), commonly tracked information about the client (*e.g.*, IP address, Autonomous System Number (ASN), etc.), and additional signals produced within Instagram, we can identify the actions initiated by each AAS. The signals produced by Instagram identify abusive services, including the AASs we study during the time of our measurement. While Instagram believes that their signals accurately characterize the entire activity of an AAS, we do not have a way to verify completeness and, as such, the levels of abuse we characterize in this section constitute a lower bound. Throughout our study, though, we *never* detect any changes in the signals tracked by Instagram for our honeypot accounts. We also periodically register additional trial honeypot accounts in each AAS as another method for observing the tracked account signals; these signals are consistent with our original honeypot accounts and also do not change during the course of our study (we delete these accounts immediately after the AAS starts generating activity on them).

In this section we analyze every action that takes place on Instagram originating from the AASs we study over a 90-day period in late 2017. This rich data set allows us to characterize the magnitude of abuse and revenue generated from AASs. We also present the types of actions performed by each service, as well as the users targeted by these actions to understand how different AASs select their targets.

Note that, for the remainder of the paper, we combine activity from Instalex and Instazood since we cannot differentiate actions performed by individual franchises (Section 3.3). To minimize confusion, we refer to their combined activity as "Insta*". Additionally, we exclude Followersgratis from the remaining analyses as the service was already well-policed by pre-existing abuse detection systems that prevent high volumes of abuse originating from a small number of IP addresses. As a result, activity generated by Followersgratis has very limited impact on Instagram in practice.

5.1 Customer Base

We explore a range of account-based measurements that help us better understand AAS operating characteristics.

Service	Customers	Long-term	Short-term
Insta*	121,661	41,891 (34%)	79,770 (66%)
Boostgram	11,959	3,975 (33%)	7,984 (67%)
Hublaagram	1,008,127	501,428 (50%)	506,699 (50%)

Table 6: Customers participating in each AAS over a 90-day period. Long-term customers of Reciprocity Abuse AASs are active beyond a trial period, and long-term Collusion Network AAS customers request service for more than four days.

Service	Operating Country	ASN Location
Insta*	Russia	USA
Boostgram	United States	USA
Hublaagram	Indonesia	GBR, USA

Table 7: The operating location for each AAS as reported on their Web site and the ASNs from service activity originates.

Popularity. How popular are these services? Table 6 shows the number of Instagram users who were active in each AAS during our measurement period. Demand for these services is large: Boostgram has more than 10,000 users, Insta* an order of magnitude more, and Hublaagram just over a million. One explanation for Hublaagram's much larger popularity is that it offers prolonged free features compared to the other AASs, and users naturally prefer no-fee services.

Since nothing constrains users from engaging with multiple services, we looked at how many Instagram users enroll their account in more than one service. Overall, account overlap is small. Fewer than 200 accounts generate any activity in the three AASs, 1,963 participate in two distinct Reciprocity Abuse AASs, and 4,485 accounts participate in at least one Reciprocity Abuse AAS as well as the Hublaagram collusion network. In these cases, nearly all are users experimenting with free trials (fewer than 100 accounts are long-term customers of any AAS).

Table 6 also breaks down the active customers into short-term and long-term categories. For Insta* and Boostgram — both of which rely on reciprocity — we define long-term users as those who participate for more than seven consecutive days, strictly longer than the length of the free trial period (Section 4.2).[2] For Hublaagram, the collusion network, we define long-term users as those who request service for more than four consecutive days. All other users are considered short-term users who only briefly engage with the services and then disappear.

One third of customers of both Insta* and Boostgram are long-term, while nearly half of Hublaagram users are long-term. Having a significant fraction of long-term uses is not surprising since, again, they offer extended services without a fee. And by far most of the actions attempted by the services come from long-term users. For Insta* and Boostgram, 91.6% and 89.7% of actions are from long-term users, and for Hublaagram it is 92.3%.

User Stability. Are AASs growing in popularity over time, or does the market appear to be saturated? Over the course of three months, we examine the rate at which new long-term users appear in each service (birth rate), the rate at which long-term users appear to have dropped out of the service (death rate), and the daily number of active long-term users in each service. Both Boostgram and Hublaagram shrank slightly over our measurement period, losing a small percentage of long-term users over time (death rate slightly higher than birth rate). In contrast, Insta* grew in size by more than 10% and the number of active long-term users per day steadily increased over the period.

Similarly, we measure the probability that a new AAS user will become a long-term user within the month they begin service. We find the long-term user conversion rate in the first month of service to be stable across our measurement period for each AAS, although the rates vary across services: the conversion rate for Boostgram is 12%, Insta* is 21%, and Hublaagram is 37%. It is not surprising that Boostgram has the lowest new long-term user conversion rate since they have the most expensive service (Table 2).

Service and Customer Location Where are customers geographically located? For each AASs we compare the country location of the service with the location of its customers. We determine the location of a service using geographic information reported on its Web site and the ASNs from which service activity originates. We define the location of an Instagram account to be the most frequent country used to login to the account, as determined by Instagram's IP geolocation system.[3]

Table 7 shows the locations of each AAS, and Figure 2 shows the countries that account for 5% or more of the user population. For each AAS, the advertised country is also where the largest number of Instagram accounts are located. Insta* has most of their users in the "other" category, which we suspect is an artifact of undiscovered franchised services around the world (Section 3.3).

5.2 Revenue

To estimate the gross monthly revenue of each service we classify the accounts participating in each service into free and paid accounts.

For Reciprocity Abuse AASs we know the account is paid when it is active in the AAS for longer than the trial period (Section 3.3.2). For each paid account we estimate the amount of money paid to the service by measuring the number of days the account is active beyond a trial period, and use the minimum paid duration as a way to convert the number of days active into money paid to the AAS. For Insta* we provide an estimated revenue range as each service (Instalex and Instazood) has a different cost and minimum service duration even though they are franchises of the same company. Table 8 shows our estimate of the monthly gross revenue for Reciprocity Abuse AASs. On average each service has a significant gross revenue approaching $200,000 to $300,000 per month.

For the collusion network Hublaagram, distinguishing between free and paid accounts is more challenging and requires a more detailed accounting methodology. Since customers can request free

[2] If an Insta* customer pays for exactly seven days of service but does not use the free trial in our measurement period, then our methodology incorrectly labels the customer as a short-term account. We expect such behavior to be infrequent, though.

[3] Note that, while AASs might affect their customer's geolocation by logging in to their Instagram accounts, they do so infrequently.

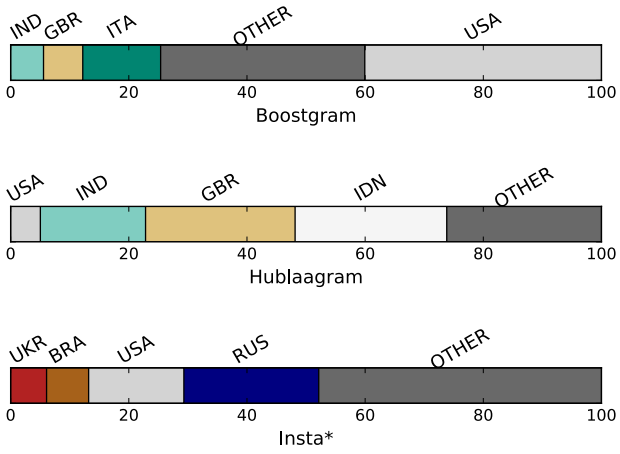

Figure 2: Percentage of AAS customer Instagram account locations by country. "OTHER" includes all countries that contribute less than 5% to the total distribution.

Service	Accounts	Service Fee	Revenue
Boostgram	3,016	$99/month	$298,584
Insta* (Low)	25,122	$0.34/day	$195,017
Insta* (High)	25,122	$3.15/week	$223,785

Table 8: Estimated monthly gross revenue for Reciprocity Abuse AASs.

service for an unbounded number of days, we cannot distinguish between free or paid solely based on the number of days they are active as we could with the other AASs. Instead, to estimate Hublaagram's monthly gross revenue we developed a model tailored to their cost structure (Table 3).

To identify accounts that pay a one-time fee to not participate in the collusion network, we count those accounts that only receive inbound actions from Hublaagram and never produce outbound actions from the service. In our measurement period, 24,420 active accounts paid the one-time fee to prevent their accounts from being used in the collusion network.

There are multiple `like` services offered by Hublaagram. To identify paying customers, for each user we count the hourly median number of `likes` generated by Hublaagram across each photo on the customer's account. Using observations from paid honeypot accounts (Section 4.1) in Hublaagram, we know that paid customers exceed the 160 `likes`/hour rate-limit imposed by Hublaagram for free customers. Therefore, we count accounts that have ever received more than 160 `likes` in an hour on any of their photos as paid since they must have purchased one of the `like` services.

For accounts classified as paid, we then distinguish among the one-time and monthly `like` services. To identify customers that purchase one-time `likes` for a single photo, we count the number of photos that have more than 2,000 `likes` for accounts that have a daily median of fewer than 250 likes per photo. Similarly, to identify customers that pay for monthly `like` services, we count accounts that have a median value of `likes`/photo that fall within the various tiers of Hublaagram's service options (e.g., we estimate an account with a median `likes`/photo ratio in the 250–500 range to be paying $20/month). We identify just 182 users who paid for one-time `likes`, while 31,901 paid for one of the monthly `like` services.

Lastly, when a customer visits Hublaagram's Web site to request free actions, they may be shown multiple advertisements that generate additional revenue for the service. The site publishes pop-under advertisements[4] from the PopAds network [25]. To increase their ad revenue, Hublaagram's Web site occasionally shows visitors pop-under advertisements on *every* Web site interaction (*e.g.*, clicking on a radio button triggers an advertisement in a new window).[5] Hublaagram provides ≈ 40 `follows` or ≈ 80 `likes` per free service request, limited to two requests per hour. We estimate the number of advertisement impressions by counting multiples of 40 `follows` or 80 `likes` performed by Hublaagram. We conservatively exclude paying customer accounts in this analysis as we are unable to differentiate paid or free `like` actions, and assume that for each request only a single advertisement was shown since we do not know how the customer interacts with the Web site. Based on PopAd's revenue model, we estimate that for every 1,000 impressions (CPM) Hublaagram receives between $0.60 and $4.00 since their customers are located around the world (Figure 2) and geolocation affects CPM [2, 6, 25].

Table 9 lists the number of paid Hublaagram accounts in each of the service categories and their contribution to overall Hublaagram's revenue.[6] Considering Hublaagram's large user base, the fraction of paid users is small. While Hublaagram had over a million active users within the measurement period, and half of them were long-term users, only about 5% of users paid fees for some kind of service beyond the free options that Hublaagram offers. Even so, Hublaagram still has an impressive estimated gross revenue of well over $800,000 per month. Most of Hublaagram's monthly revenue derives from customers paying for 250–1,000 `likes`/photo per month, while few customers purchase one-time `likes` for a single photo (reflecting how poor a bargain that option is). Similarly, while many ads are shown, we estimate that the resulting ad revenue is dwarfed by the other revenue sources.

Interestingly, users *do* care about not receiving fake outbound actions from other accounts in the collusion network, and are willing to pay for preventing it. Of the active accounts in our observation period, such users collectively paid Hublaagram more than $350,000 in one-time fees.

A related question is if the majority of monthly AAS revenue is generated from customers that pay for service only once, or ones that renew. Table 10 shows the fraction of new paid customers versus customers that have paid for service before. Across all services, the majority of gross revenue is generated from AAS customers who repeatedly pay for service.

[4] Pop-under ads typically appear when closing a Web page.
[5] Hublaagram's Web site shows between 1–4 pop-under ads per free service request.
[6] Fewer than 20 customers mapped to the 5,000 or 10,000 one-time `like` service categories, and we exclude them from Table 9 since their revenue contribution is negligible.

Service	Accounts	Fee	Revenue
No outbound	24,420	$15	$366,300
Total One-Time Revenue			$366,300

Service	Count	Fee	Revenue
Ads Shown			
Low CPM	5,769,537	0.06¢	$3,461
High CPM	5,769,537	0.4¢	$23,078
Likes Once			
2,000	182	$10	$1,820
Likes / Photo			
250 − 500	11,249	$20	$224,980
500 − 1,000	18,009	$30	$540,270
1,000 − 2,000	2,488	$40	$99,520
2,000 − 4,000	155	$70	$10,850
Total Monthly Revenue			$880,901 – $900,518

Table 9: Gross revenue estimates for Hublaagram. The "No outbound" service has a one-time fee for the lifetime of the account, and the remaining services have monthly fees.

Service	New	Preexisting
Insta*	31.4%	68.6%
Boostgram	10.8%	89.2%
Hublaagram	16.5%	83.5%

Table 10: Breakdown of revenue between new and existing paying customers for each AAS over one month.

5.3 Activity Generated

We now analyze the *actions* performed by each AAS to understand which types are most popular among users, and how Reciprocity Abuse AASs target specific kinds of users to obtain better organic reciprocation rates.

Table 11 shows the proportion of action types performed by each AAS throughout the measurement period. `Likes` are the most requested action for Boostgram and Hublaagram, 1.8–3.4× more popular than `follows`. Insta* customers request more `follows` to `likes` (1.3×). Across all AASs, `comments` and `posts` are infrequent, suggesting that customers of these AAS either acquire these actions through other means, or do not consider them as valuable. The Reciprocity AASs perform a significant number of `unfollows`, which users can optionally request to happen automatically after a `follow`.

Reciprocity AASs depend upon general Instagram users to generate reciprocating `follows` and `likes` to their customers' requests. As a result, if these services can target Instagram users who are more likely to reciprocate, they can more easily meet their customer demands. To evaluate whether Reciprocity Abuse AASs have any biases in the accounts that they target, we compare accounts targeted by actions from AASs with accounts from all of Instagram as a baseline. Specifically, we compare the following and follower counts of a random sample of 1,000 accounts that received an action from AASs with a random sample of 1,000 from all Instagram accounts that receive actions during our measurement period.

Action	Insta*	Boostgram	Hublaagram
Likes	30.8%	64.0%	63.0%
Follows	38.6%	19.3%	35.3%
Comments	5.6%	0%	1.7%
Unfollows	25.0%	16.7%	0%

Table 11: Action types performed from each AAS over a 90-day period. We normalize each value by the total number actions performed by each service.

For both metrics we see differences in the account populations. Figure 3 shows a CDF of the number of Instagram accounts followed by the accounts in each sample (account out-degree). For example, the median AAS accounts have a higher out-degree than a random Instagram account: Boostgram accounts follow 684 other Instagram accounts and Insta* accounts follow 554.5, while the median sample of all of Instagram accounts follow just 465. Similarly, Figure 4 shows a CDF of the number of followers of the accounts in each sample (account in-degree). By this metric, the distributions have even more pronounced differences: The accounts targeted by the Reciprocity AASs have significantly fewer followers than the broader Instagram population. Boostgram and Insta* accounts are followed by just a median of 498 and 384 accounts, respectively, whereas the median sample of all Instagram accounts are followed by 796 accounts.

These results indicate that the Reciprocity AASs do have a selection bias in the accounts that they target, selecting for accounts with higher out-degree and much lower in-degree to increase the likelihood of a reciprocated action. In other words, accounts targeted by the AASs are already inclined to follow other users, but have far fewer followers themselves and, as a result, are presumably more open to reciprocating when targeted.

6 INTERVENTIONS

Having characterized AAS from a user perspective and as business entities, we subsequently actively engage with the abusive services by deploying countermeasures. Our goal is not to completely disrupt the AASs immediately, but rather we start by evaluating how AASs react to interventions. This understanding can then provide insight for improving operational abuse detection and prevention systems. While Instagram is in a position to identify all AAS customer accounts, blocking these accounts is not a desirable outcome since Instagram users still use them to initiate legitimate actions that should not be blocked (even while they are also enrolled in an AAS). Additionally, as our interventions show in Section 6.3, AASs quickly attempt to evade interventions. As such, we derive a new signal for performing countermeasures (Section 6.2), rather than relying on the signals used to identify AAS customers in the first place. We perform two interventions, first on a narrow set of AAS activity over a six-week period, and a second on a broad set of AAS activity over a subsequent two-week period.

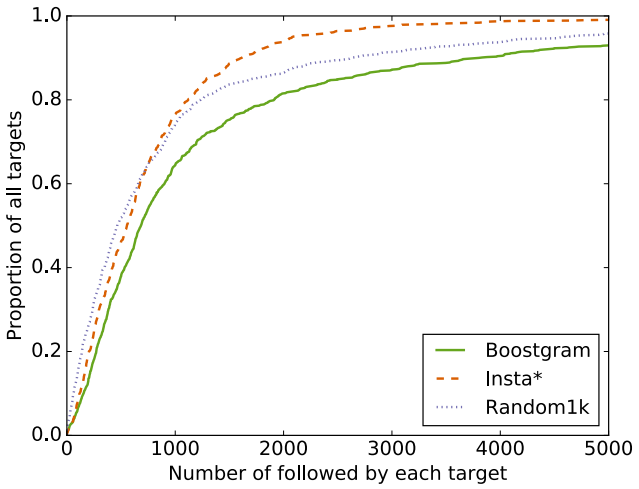

Figure 3: CDFs of the number of users followed by each target for three samples of accounts: $1,000$ random accounts targeted by the two Reciprocity AASs, and $1,000$ random Instagram users.

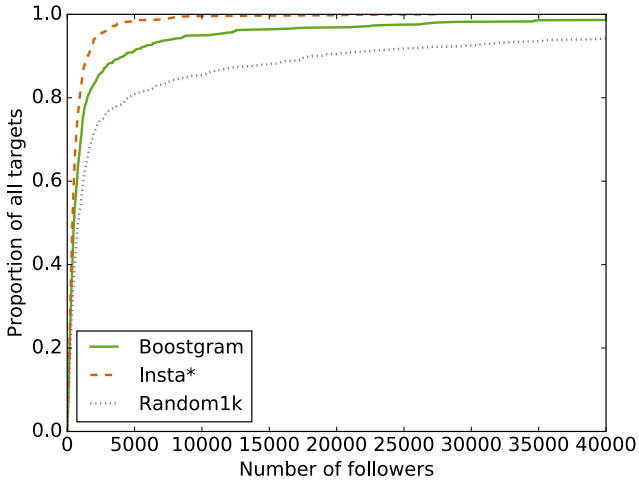

Figure 4: CDFs of the number of followers for a random sample of $1,000$ targets selected by two third-party applications compared to a sample of $1,000$ Instagram users.

6.1 Countermeasures

Instagram has a variety of options for reducing or disrupting the impact of an abusive action, and we experiment with two. Each countermeasure response comes with a trade-off between its effectiveness at disrupting abuse, and the ease with which an adversary detects the intervention.

Synchronous Block. When blocking AAS actions, the actions are not successful and do not reach users of Instagram. Such a countermeasure directly undermines the perceived value of using an AAS. At the same time, though, the transparent aspect of the synchronous response serves as an oracle of what actions Instagram can detect as abusive. The AAS can use this oracle to easily test and possibly adjust their strategy for delivering their actions to accommodate or sidestep the countermeasure within a short period of time.

Delayed Removal of Follows. With the delayed removal countermeasure, `follows` from accounts used by AASs are initially successful but then are removed by Instagram one day after taking place. The deferred nature of the delayed response helps mask the countermeasure as it is more difficult for AASs to realize their actions are being detected. Note that we only apply this countermeasure to `follow` actions, as it was not possible to apply a delayed countermeasure on `likes`.

6.2 Identifying Eligible Actions

As with all anti-abuse measures, from spam filtering to anti-virus, one must balance the value provided in addressing abusive behavior against the unintentional misclassification of a benign action. Thus, while AASs are insidious in undermining the confidence in the integrity of the content being posted, so too must we consider and be sensitive to users whose legitimate actions might be inadvertently blocked or removed. To this end, we have carefully designed our interventions to minimize these risks; throughout the duration of our experiments we identified a handful of false positives and these were remediated manually.

In particular, we start by focusing on actions from the small number of ASNs that the AASs use. Then we define a per-account daily activity threshold for each ASN, and only actions above that threshold are candidates for a countermeasure. The threshold is defined in terms of legitimate activity, so activity by an account above the threshold strongly suggests abusive behavior by that account. Specifically, we track the number of outbound actions from Instagram accounts used by the Reciprocity Abuse AASs, and we track the number of inbound actions from accounts used by the Collusion Network AAS. We use the same methodology from Section 5 combined with paid honeypot accounts to track AAS activity and reactions to countermeasures.

Note that we compute the activity thresholds differently across ASNs since some ASNs have only AAS traffic while others have benign user activity blended in. For ASNs with both AAS and benign traffic, we measure the daily 99th percentile of `likes` and `follows` produced by Instagram accounts that are not participating in AASs. Since accounts involved in AASs produce significantly more actions than non-AAS accounts, using the daily 99th percentile of non-AAS activity represents an upper bound of 1% false positives. For ASNs with only AAS traffic, we use a threshold of the daily 25th percentile of actions since there is no legitimate user traffic from those ASNs.

We computed the activity level thresholds at the start of each experiment and did not change them to prevent an adversary from affecting the false positive rate. Throughout both experiments we actively monitored complaints to Instagram from users who *might* be affected by our experiments. We received only a handful of complaints from legitimate users who were inadvertently impacted which we worked to address. In contrast, we also monitored complaints to the AASs from their customers, and some of the interventions generated highly voluble complaints.

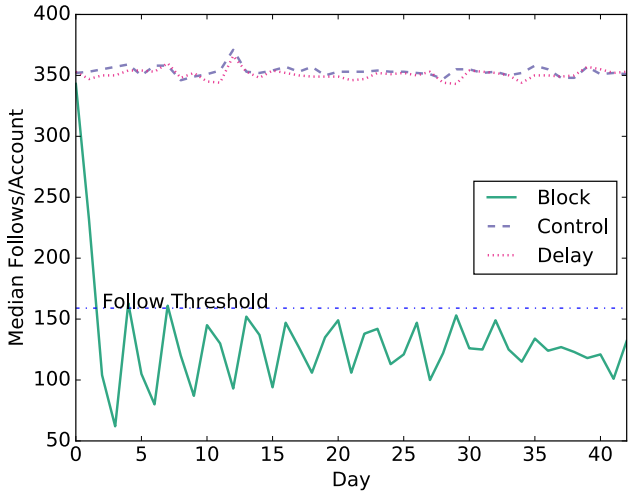

Figure 5: Median follows per user each day participating in Boostgram. We show the countermeasure threshold as a dashed line, and the median actions for both users who are blocked by countermeasures, and in our control (no countermeasures)

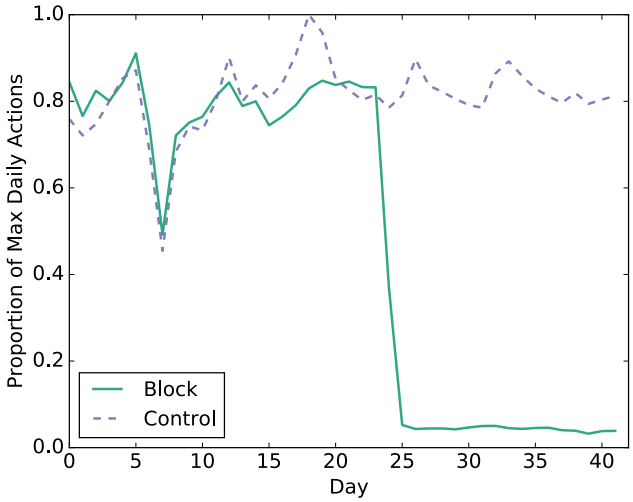

Figure 6: The proportion of Hublaagram likes each day that are eligible for a countermeasure. We noticed at around the third week the service makes a strict adjustment significantly reducing the number of eligible likes.

6.3 Narrow Interventions

In our first intervention we evaluate how AASs react to the countermeasures from Section 6.1 when they are continuously applied for six weeks to the same subsets of AAS customers. To define different sets of Instagram accounts that may receive a countermeasure response, we deterministically partition Instagram accounts into 10 equally-sized bins. We assign separate bins for each countermeasure response (block and delay) and another for a control. By partitioning Instagram accounts into 10 bins, each bin contains at least 5% of long-term customers (for each AAS) that produce actions eligible for a countermeasure (Section 6.2). Throughout a six-week period in 2017, we continuously apply each of the two countermeasure responses to all eligible AAS actions that go above the daily activity threshold when the Instagram account is within a particular countermeasure bin. Accounts in the control bin never receive a countermeasure even when actions go beyond the activity threshold. In total, this experiment applies countermeasures to at most 20% of the customers in each AAS.

When applying the countermeasures to `follow` actions, all of the AASs react similarly. Figure 5 shows Boostgram activity as a representative example. Each curve shows the median number of actions per Instagram account in each countermeasure bin and the control bin for each day of the six-week period of the experiment. The dashed "Follow Threshold" line shows the threshold above which the countermeasure affects actions in Instagram. The service reacts immediately to blocking `follows`, dropping the number of actions below the threshold and probing it thereafter. Boostgram (and the other services) clearly detect that blocking is taking place, and the reaction patterns across services strongly suggests that it is an automated process; indeed, we found an openly available implementation of one of these services with block detection logic. Countermeasures that provide a strong signal to the services unfortunately enable them to adapt, and adapt quickly.

Even more interesting, though, is that the services *do not* react to delayed removal of `follows`, even though the countermeasure undoes all of the activity one day later. Ironically, delayed forms of countermeasure satisfy both sides: the services successfully perform `follows` and continue on apparently unaware that the countermeasure cleans them up shortly afterwards as if they never happened. (Customers of the services, though, lose out.) Blocking and delayed removal both ultimately have the same benefit to Instagram—`follow` actions are truncated to the threshold — but blocking provides a signal to services, while delays do not.

Only Hublaagram reacts when we apply the countermeasures to `likes`, presumably since `likes` are its primary source of income. Figure 6 shows the proportion of daily `likes` above the threshold that the countermeasures can affect. Again, Hublaagram only reacts to blocking and, because blocking is straightforward to detect, it is able to drop its `like` activity and discover the threshold under which blocking does not take place. Hublaagram does take three weeks into the intervention period to react, perhaps because it had to implement blocked `like` detection.

6.4 Broad Interventions

Our first intervention applied each countermeasure to a narrow 10% of users, perhaps so narrow that the services did not fully notice or react to countermeasures (delay removal in particular). Consequently, our second intervention applied the delay and block countermeasures broadly to 90% of the AAS user accounts, keeping the same 10% bin of control accounts as before. In this experiment we apply the delayed removal for one week, and then blocking for another.

The reactions of the AASs to the broad intervention are similar to their reactions for the narrow intervention. As representative behavior, Figure 7 shows the proportion of daily Boostgram `follows` above the activity threshold that are subject to countermeasures. The control bin, with 10% of accounts, appropriately has 10% of the

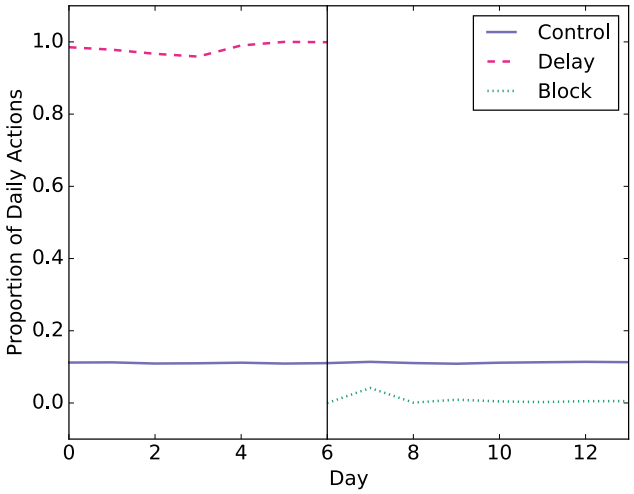

Figure 7: Proportion of Boostgram follows eligible for countermeasures each week during the experiment. On day 6, we switched the countermeasure response from delay to block, shown by a vertical line.

actions above the threshold throughout. In the first week we deploy the delay countermeasures to the remaining 90% of accounts, again with no reaction by Boostgram — even though the countermeasure now applies to actions above the threshold for nearly all of their users. We then replace delay with the block countermeasure for the second week. As with the narrow intervention, Boostgram detects that their `follows` are being blocked and scales back their actions to the threshold.

Epilogue. The broad intervention remained active, continuing to block `likes` and delay `follows` above the activity threshold for additional months. Since the services immediately detected blocked actions, all AASs eventually moved their `like` traffic to different ASNs — one of them going so far as to use an extensive proxy network to drastically increase IP diversity. As a result, the `like` actions from the AAS were subsequently out of reach of the blocking countermeasure we employed, underscoring the risks of a countermeasure so easily detected.

After a few months, Hublaagram, unable to produce sustainable unblocked actions, stopped accepting customer payments by listing all offered services as "out of stock". Insta*, on the other hand, eventually moved their `follow` actions back into the original ASN in which we applied the delayed intervention.

7 CONCLUSION

Social networks such as Instagram attract abuse because they provide a mechanism for attracting and focusing the attention of large groups. Whether for social or economic reasons, a range of users are interested in artificially inflating their standing in such networks — paying to acquire thousands of `follows`, pervasive `likes` of their photos and so on. Simplistic approaches to manipulate social standing (i.e., using fake accounts) can be readily detected and thus sophisticated services have emerged that remotely "drive" the accounts of their customers to manipulate their social standing in a manner more likely to appear organic. We have identified two common techniques used to achieve this end on the Instagram network — driving outbound `follows` to attract reciprocal `follows` (reciprocity abuse) and laundering social actions across a network of customer participants (collusion networks). We've shown that services using these techniques have been successful in attracting and maintaining long-term customers generating per-service revenues between $200k-900k per month. Finally, we have shown through controlled experiments that blocking such services, while effective in the short term, quickly drives adaptation and can make it difficult to amortize the cost of developing accurate abuse classification. Consequently, from the standpoint of protecting non-abusive users from artificial content, a more effective long-term strategy can be built on deferred interventions (e.g., removing synthetic actions after at a future point). Such approaches greatly increase the "debug time" for services seeking to reverse engineer how they are being detected and are less likely to drive the customer complaints that incentive services to pursue such adaptations.

ACKNOWLEDGEMENTS

We thank our shepherd Gianluca Stringhini and the anonymous reviewers for their insightful feedback and suggestions. This work was supported in part by NSF grants CNS-1629973 and CNS-1705050, DHS grant AFRL-FA8750-18-2-0087, the Irwin Mark and Joan Klein Jacobs Chair in Information and Computer Science, and by generous research, operational and/or in-kind support via the UCSD Center for Networked Systems (CNS).

REFERENCES

[1] AGGARWAL, A., AND KUMARAGURU, P. What They Do in Shadows: Twitter Underground Follower Market. In *Proceedings of the 13th Conference on Privacy, Security and Trust (PST)* (Izmir, Turkey, July 2015).
[2] BLOGNIFE. PopAds CPM Rates 2018. http://blognife.com/2017/06/22/popads-cpm-rates-2017/, 2017.
[3] BOOSTGRAM. Boostgram Web site. https://boostgram.com, 2017.
[4] CHAUM, D. Untraceable Electronic Mail, Return Addresses, and Digital Pseudonyms. *Communications of the ACM* (February 1981), 84–88.
[5] DE CRISTOFARO, E., FRIEDMAN, A., JOURJON, G., KAAFAR, M. A., AND SHAFIQ, M. Z. Paying for Likes? Understanding Facebook Like Fraud Using Honeypots. In *Proceedings of the ACM Internet Measurement Conference (IMC)* (Vancouver, BC, Canada, November 2014), pp. 129–136.
[6] EARNING GUYS. PopAds Review: A Pop-under Ad Network. http://www.earningguys.com/advertisement/popads-review/, 2017.
[7] FAROOQI, S., ZAFFAR, F., LEONTIADIS, N., AND SHAFIQ, Z. Measuring and Mitigating Oauth Access Token Abuse by Collusion Networks. In *Proceedings of the ACM Internet Measurement Conference (IMC)* (London, UK, November 2017), pp. 355–368.
[8] FOLLOWERSGRATIS. Followersgratis Web site. http://followersgratis.org, 2017.
[9] FSTOPPERS. Mass Planner Shut Down by Instagram: The End of the Bot Era. https://fstoppers.com/social-media/mass-planner-shut-down-instagram-end-bot-era-176654, 2017.
[10] HOOI, B., SONG, H. A., BEUTEL, A., SHAH, N., SHIN, K., AND FALOUTSOS, C. FRAUDAR: Bounding Graph Fraud in the Face of Camouflage. In *Proceedings of the 22nd ACM International Conference on Knowledge Discovery and Data Mining (KDD)* (San Francisco, CA, USA, August 2016), pp. 895–904.
[11] HUBLAAGRAM. Hublaagram Web site. http://hublaagram.me, 2017.
[12] INSTAGRAM. Strengthening Our Commitment to Safety and Kindness for 800 Million. http://blog.instagram.com/post/165759350412/170926-news, 2017.
[13] INSTAGRAM. Terms of Use. https://help.instagram.com/581066165581870, 2018.
[14] INSTALEX. Instalex Web site. https://instalex.ru, 2017.
[15] INSTALEX FRANCHISE. Instalex Franchise Web site. https://instalex.pro/franchise, 2017.
[16] INSTAZOOD. What is a Good Engagement Rate on Instagram. https://instazood.com/what-is-a-good-engagement-rate-on-instagram/, 2017.
[17] INSTZOOD. Instzood Web site. https://instazood.com, 2017.
[18] JAVED, M., HERLEY, C., PEINADO, M., AND PAXSON, V. Measurement and Analysis of Traffic Exchange Services. In *Proceedings of the ACM Internet*

Measurement Conference (IMC) (Tokyo, Japan, October 2015), pp. 1–12.

[19] LEE, K., CAVERLEE, J., AND WEBB, S. Uncovering Social Spammers: Social Honeypots + Machine Learning. In *Proceedings of the 33rd ACM Conference on Research and Development in Information Retrieval (SIGIR)* (Geneva, Switzerland, July 2010), pp. 435–442.

[20] MEDIUM. Instag-RAMPAGE: the WAR on Automation. https://medium.com/@mountainbeard/instag-rampage-and-the- war-on-automation-3a7362b08112, 2017.

[21] MEDIUM, SHANE BARKER. How to Become an Instagram Influencer and Start Earning Money Now. https://medium.com/swlh/how-to-become-an-instagram-influencer-and-start-earning-money-now-a8ef3169e96d, 2018.

[22] MISLOVE, A., MARCON, M., GUMMADI, K. P., DRUSCHEL, P., AND BHATTACHARJEE, B. Measurement and Analysis of Online Social Networks. In *Proceedings of the ACM Internet Measurement Conference (IMC)* (San Diego, CA, USA, October 2007), pp. 29–42.

[23] NEW YORK TIMES. How Bots Are Inflating Instagram Egos. https://www.nytimes.com/2017/06/06/business/media/instagram-bots.html, 2017.

[24] NEW YORK TIMES. The Follower Factory. https://www.nytimes.com/interactive/2018/01/27/technology/social-media-bots.html, 2018.

[25] POP ADS. PopAds Web site. https://www.popads.net, 2017.

[26] STRINGHINI, G., KRUEGEL, C., AND VIGNA, G. Detecting Spammers on Social Networks. In *Proceedings of the 26th Annual Computer Security Applications Conference (ACSAC)* (Austin, TX, USA, 2010).

[27] STRINGHINI, G., WANG, G., EGELE, M., KRUEGEL, C., VIGNA, G., ZHENG, H., AND ZHAO, B. Y. Follow the Green: Growth and Dynamics in Twitter Follower Markets. In *Proceedings of the ACM Internet Measurement Conference (IMC)* (Barcelona, Spain, October 2013), pp. 163–176.

[28] THE VERGE. Popular Instagram bot site Instagress has been shut down. https://www.theverge.com/2017/4/20/15374080/instagram-bot-site-instagress-dead, 2017.

[29] VISWANATH, B., BASHIR, M. A., CROVELLA, M., GUHA, S., GUMMADI, K. P., KRISHNAMURTHY, B., AND MISLOVE, A. Towards Detecting Anomalous User Behavior in Online Social Networks. In *Proceedings of the 23rd USENIX Security Symposium* (San Diego, CA, USA, August 2014), pp. 223–238.

[30] WEBB, S., CAVERLEE, J., AND PU, C. Social Honeypots: Making Friends With A Spammer Near You. In *Proceedings of the Fifth Conference on Email and Anti-Spam (CEAS)* (Mountain View, CA, USA, August 2008).