DOMAINTOOLS®

Spring 2021

THE DOMAINTOOLS REPORT

Domain Blooms: New Method of Detecting Trending Bad Domains

Every domain on the Internet has a job to do, and analysis of naming schemes of domains can often tell a researcher a lot about the domain's intended purpose. At first, it might seem that little to no analysis is required, since the meaning behind the name appears self-evident.

However, the widespread use of domains for fraud, cybercrime, information warfare, and other such objectives makes at-a-glance name-based assessment of domains unreliable. Domains registered for malicious purposes often spoof legitimate brands, persons, organizations, or topics of general interest and discussion, aiming to deceive Internet users into visiting those domains, furthering the objectives of the domains' owners. Looking at the name alone, it is often difficult or impossible to evaluate the domain's legitimacy or intended purpose.

Because most* domain names are intended to be read, interpreted, and acted upon by a human, domain naming schemes usually must have some meaning in order to be effective. In the tumultuous year that was 2020, events around the world created a vast hunger for information, goods, services, etc, tied to those events, and hundreds of thousands of domains were registered to capitalize on the demand. Many of those domains were used in good faith for legitimate, beneficial purposes, but, as is the case with the Internet in general, many were used maliciously. This report examines patterns of domain registrations in 2020 based on two significant event clusters of that year, to better understand how domain registrants leverage these events, and to help security professionals and researchers make efficient use of available data to identify "hot spots" of dangerous or questionable activity in the interest of making the Internet a safer place for everyone.

***Some names are not meant for human consumption. The most obvious category here is domains used for malware command and control (C2), where it is a bot rather than a human that establishes a connection to the domain. C2 domains are often seemingly-random strings of alphanumeric characters generated by domain generation algorithms (DGAs) and registered in large tranches. However, some DGAs do use real words, even though the domains are still meant for bots rather than people.**

Objectives of the Research

With plenty of evidence already in hand of a huge number of domain registrations tied to the COVID-19 pandemic, and having done extensive analysis in cooperation with the COVID-19 Cyber Threat Coalition (CCTC), we assumed that other significant events of 2020 would likewise give rise to large numbers of registrations. To test this assumption, we began an examination of the DomainTools database to find patterns of registrations matching keywords tied to significant events of the year, and to help identify threats among those domains.

The work originated when our research team, particularly Sean McNee, Chad Anderson, and report co-author John "Turbo" Conwell, started looking at domain names related to COVID/Coronavirus in December of 2019. As the weeks went by, the numbers of domains with keywords related to the emerging pandemic grew significantly; by the middle of March 2020, we observed thousands of new registrations per day with COVID-related terms, and their common phishing-related variants, in the names. As a security company, we pay particular attention to malicious use of online infrastructure, so much of our effort focused on sorting out the benign from the dangerous. The COVID-related domain name themes provided a laboratory of sorts in which the team developed new automated analytical processes to deal with the large scale of domain creation we were studying. The work paid off, helping the CCTC develop a COVID-themed blocklist of dangerous domains related to the pandemic.

With that work under our belts, we set out to apply the same approach to other news-making events of 2020. It is our hope that the analytics and insights we have developed will help the information security community more broadly in their efforts to identify and isolate dangerous domains and their associated infrastructure.

Methodology for This Report

We began the process simultaneously from two directions: we brainstormed a set of keywords to examine, based on a review of the most significant world events in 2020; and we also created algorithms designed to naïvely identify a pattern we call domain blooms. The dual approach turned out to be useful, because the **domain bloom** work found not only the expected patterns around some notable events of the year, but also some blooms whose naming schemes did not make immediate sense from a semantic perspective, but which did make sense on further analysis. The discussion will touch briefly on the latter subject, as it may be of use in its own right to researchers and network defenders.

It is easy to find almost any actual word (at least in a globally popular language) in the DomainTools database of domain registrations. The more challenging task is to identify changes to the frequency in which that word occurs among the hundreds of millions of domains already in existence and the hundreds of thousands that are typically registered each day.

The algorithm the research team developed compares every domain name registered each day against a curated dictionary of known words, which establishes a background rate of word occurrences per day. This dictionary is updated daily from current event news to ensure it doesn't get stagnant. By looking at registrations over discrete time windows, it is able to identify changes to this background signal over time. **The analysis uncovered two patterns of deviation from the norm, which we have called "spikes" and "blooms."**

> A **spike** is a sharp increase and an equally sharp decrease in the rate of the word's occurrence in registrations. Some spikes are as short as a single day; others may last a bit longer, but unlike blooms, they return quickly to the prior level.
>
> A **bloom** is a pattern characterized by a steep rise in the frequency of a word across domains registered within a given time period, followed by a gradual return to a steady state. For some blooms, the post-bloom state is higher than before the bloom; this would be expected when a new word, such as "COVID," has entered the global lexicon. For other blooms, the post-bloom signal looks very similar to a pre-bloom. Blooms often last weeks or even months

Why do some words appear in domain blooms, and others in spikes? Without surveying the registrants, we are left to conjecture, but some inferences are reasonable. For example, many spikes could represent the efforts of speculators, hoping that domains that were registered quickly after a notable event might prove valuable properties as the event entered the zeitgeist. Domain speculation of this nature is almost as old as the Internet itself. If the news cycle left the event behind after a day or two, the payoff became much less likely and the registrations fell off precipitously. Blooms, by contrast, suggest a combination of speculation and intentions to make (good or ill) use of words that entered—and showed a likelihood of remaining in—the common vocabulary over time. The bloom pattern clearly shows that registrants were motivated by something to keep registering domains over time. From a security point of view, both blooms and spikes could have potentially interesting implications, particularly if there are any commonalities to be observed among the malicious portion of the domains represented in these patterns.

Since identifying the "malicious portion" is important to this effort, another part of the analysis for this report involved analyzing predictions about which of the domains in the blooms or spikes were intended for malicious purposes. For the predictions, we relied on the **DomainTools Risk Score**, a product developed by Sean McNee, John "Turbo" Conwell, and Michael Klatt.

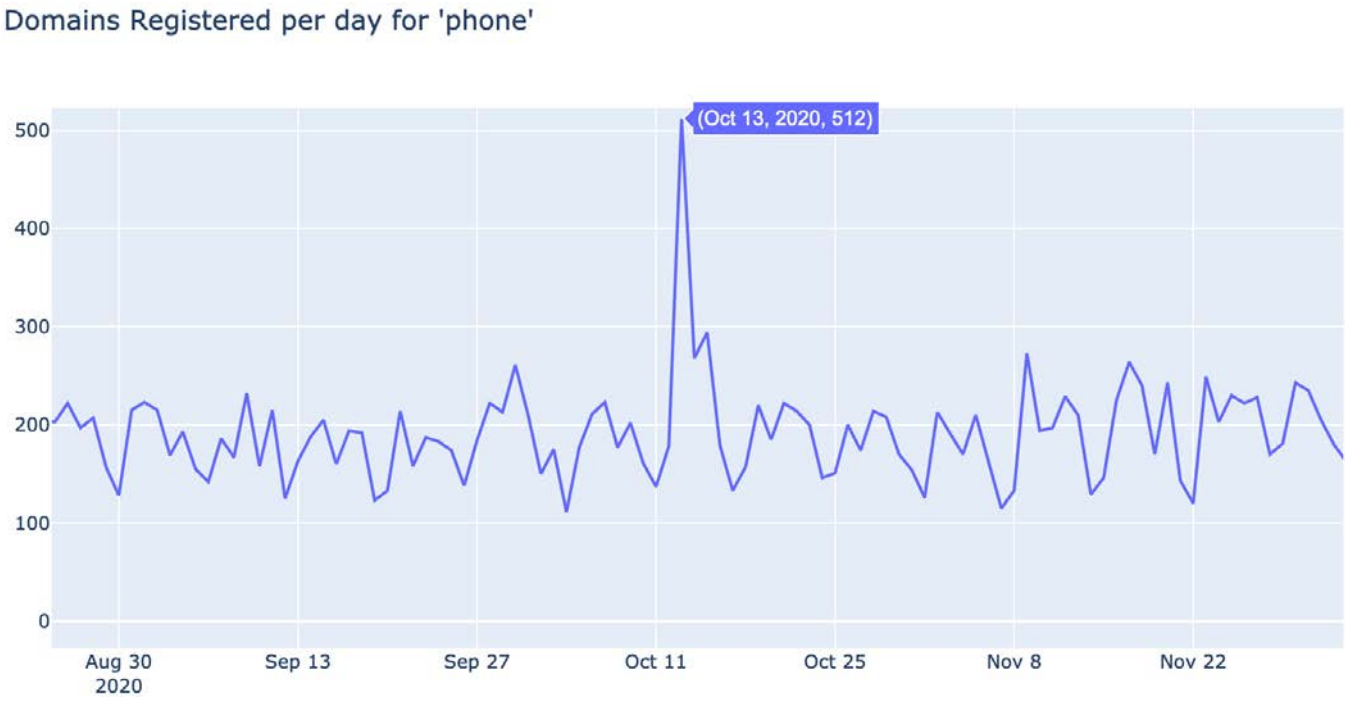
In general, we consider domains with Risk Scores over 70 to be highly suspicious. We acknowledge that no scoring system is foolproof. In general, however, and especially at the scales of the analysis for this report, the scores give a reliable profile of the relative occurrence of likely-bad domains among the overall population.

The Domain Bloom Algorithm

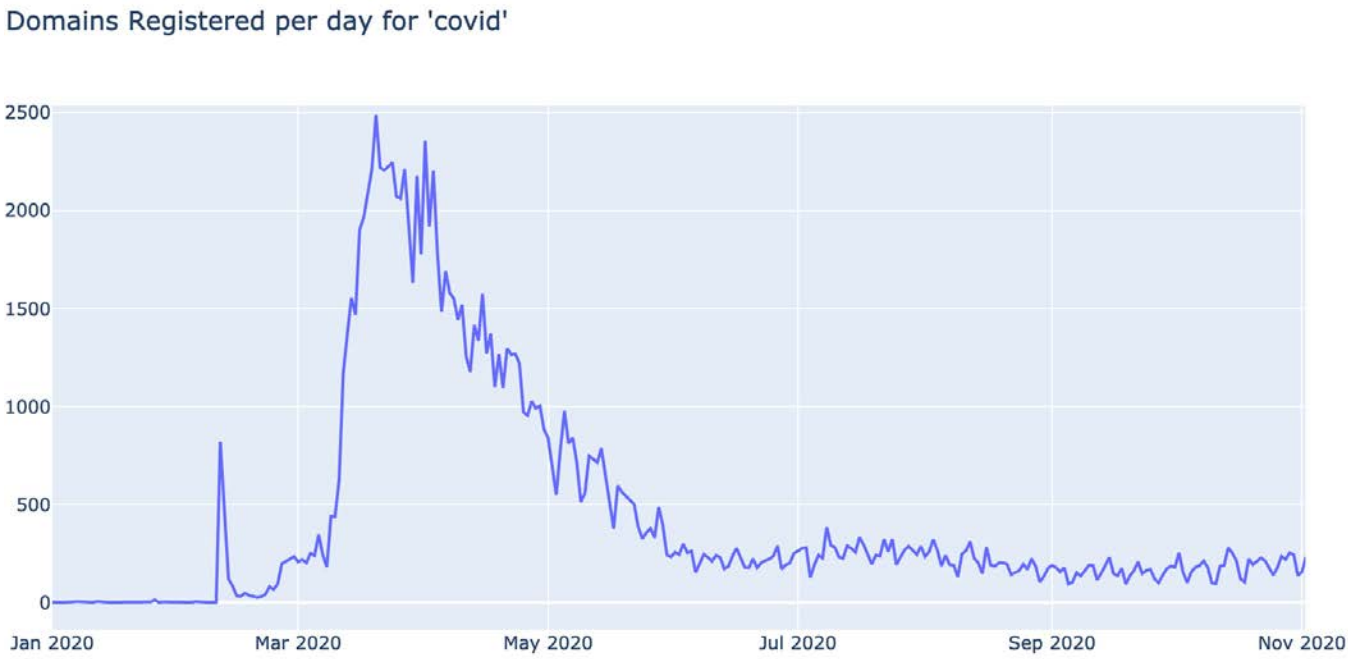
A development of work we had been doing around word frequency in domain registrations, the idea behind Domain Blooms is pretty simple: take all the domain names registered on a given day and break them into their individual component words (discarding the TLD). For example, bankofamerica[.]com would be split into the words: "bank", "of", "america", "fame". From this, we group the same words together and get their count. This gives us the frequency of unique words used to register domains for that day. Then we calculate this for every day going back in history, as well as for each new day.

With this data set we can then see how often an individual word is used to register domain names over time. Some words are rarely, if ever, used, like “COVID” before February of 2020. Others are used almost consistently every day, like “phone” shown in the following graph:

Here we can see that “phone” is used in approximately 180 domains per day. This can be viewed as the “baseline” usage frequency of the word “phone” over time. What becomes interesting is when we identify outliers from this baseline. An example is the date range below: on October 13 there was a noticeable spike in domains registered with the word “phone” that is 300 domains greater than the baseline. This was the day of the iPhone 12 launch.



To see the difference between spikes and blooms, take a look at the graph for the word “COVID” in 2020, specifically from March to July:



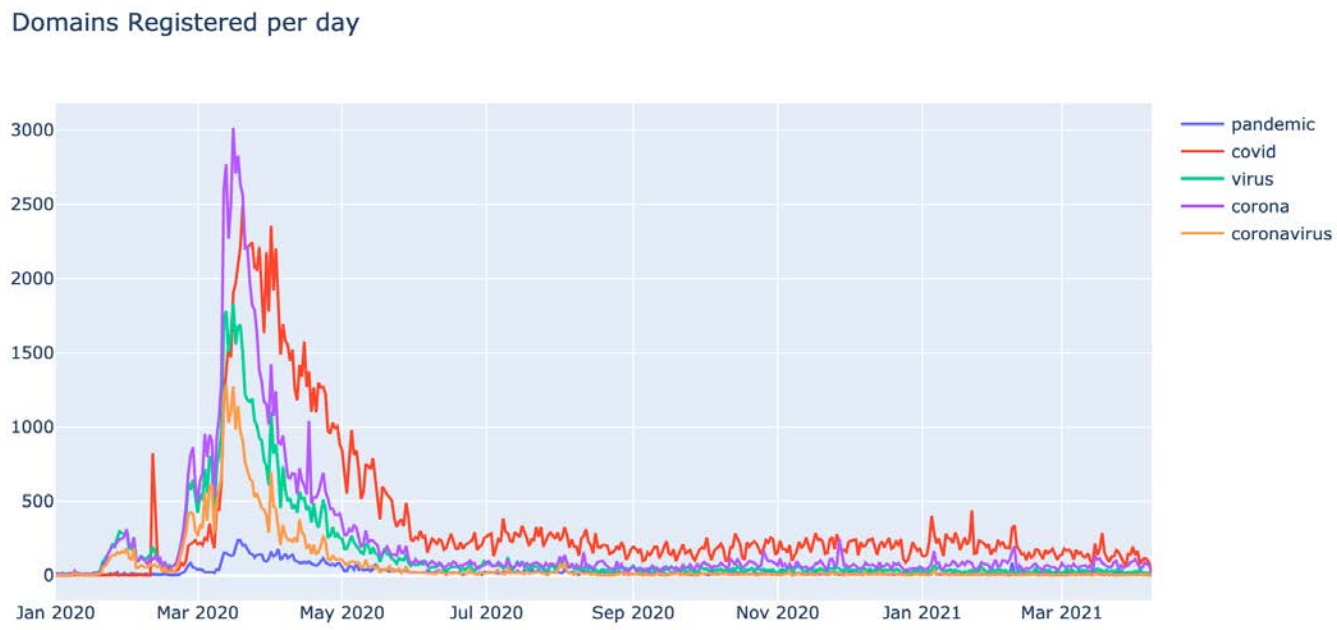
This is a perfect example of a bloom: a rapid increase from the word’s baseline, but unlike a spike, after a few days the number of domains registered per day with the given word either continues to increase, or holds steady over a period of time. After a while the number of such domains registered per day gradually drops back to its baseline level, or potentially a new baseline level that remains steady.

Note: For a more technically detailed description of the Domain Bloom algorithm, its design, and the types of analysis it enables, please see the companion blog post by Turbo Conwell.

Clustering Domain Blooms into "Bouquets"

Since a domain bloom is really just a time series histogram, or count of occurrences over time, if we take any given bloom and search the "bloom space" for other histograms that have a similar shape, we should be able to infer some relationship between blooms that are similar to each other.

To test this, we used a special type of machine learning algorithm called clustering that groups similar items together. We first generated the bloom histogram for all known words used in domains registered during 2020, and then ran all the blooms through the clustering algorithm. When we look at the top 5 words that grouped together in the "COVID" cluster, we get this plot. Within it, we can see the bloom for 'covid' as depicted earlier, along with the blooms for several other related words:



Notice that not all the blooms have the same magnitude as each other, but they have a very similar relative shape. They start around the same time, and have a similar duration. We can use these bouquets to identify other words that are being used in a similar way with domain registrations.

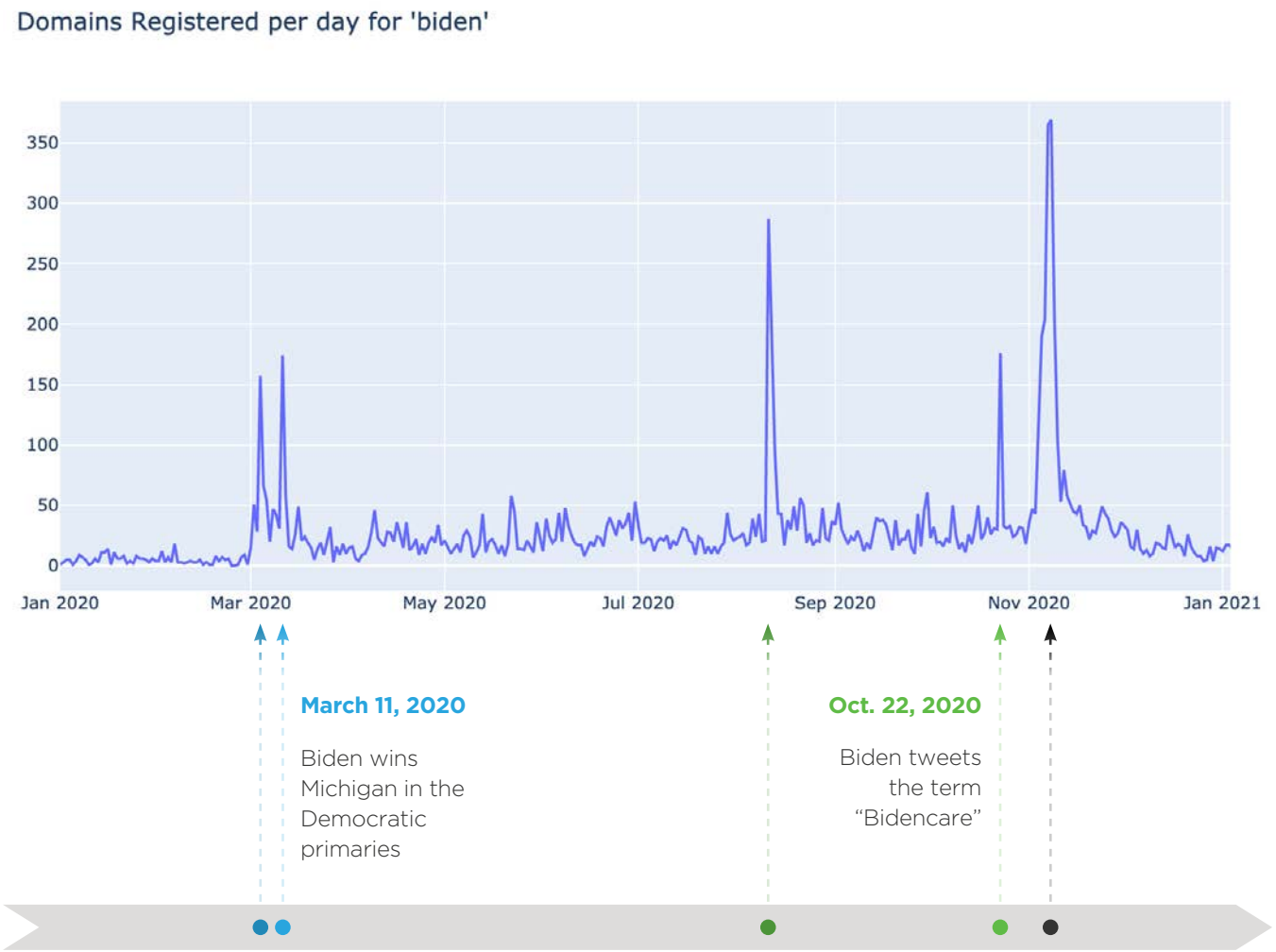
Findings

2020 Domain Spikes

Domain Name Speculation

We estimate that one of the primary reasons we see spikes is for domain name speculation. Such spikes are almost always associated with a news event, with domain registrants seemingly trying to quickly "land-grab" as many potential relevant domains as possible.

One of the best words to demonstrate domain spikes during 2020 is the word "biden". If we look at the domain registration timeline for the word "biden" we can see 5 spikes during 2020. These spikes are either on, or one day after, some major news event:



When we look into the domains that make up these spikes, we usually see a common pattern: the majority of them are parked, show little to no passive DNS activity, and typically have the same registrar and limited set of TLDs.

Below is an example of "bidencare" domains registered on October 23, 2020 that fit this profile:

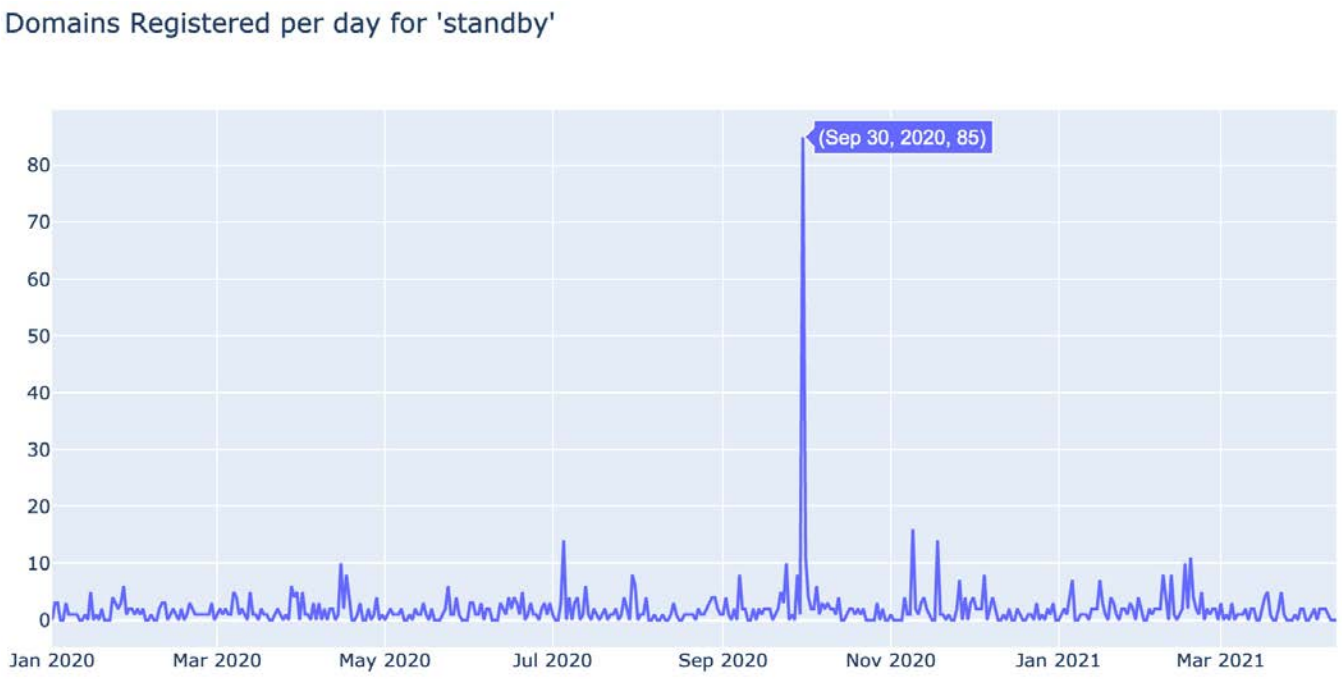
```
bidencarebenefit[.]com          bidencarequote[.]com
bidencareflorida[.]com          bidencarerates[.]com
bidencarenetwork[.]com          bidencaretexas[.]com
bidencareoptions[.]com          bidencareunity[.]com
bidencarewebsite[.]com          bidenhealthcare[.]us
bidenCOVID19plan[.]com          bideninsurance[.]com
bidenhealthcare[.]info          bideninsurance[.]org
 bidencarebuyin[.]com            bidencare[.]support
 bidencarefacts[.]com            bidencareplan[.]org
 bidencarefacts[.]org
```

The driving factor behind speculation spikes seems to be the hunch that there will be some future value in at least some of the registered domains. By taking a bit of a shotgun approach to investing, the registrant hopes that one or more of the domains will be in demand in the future and can be sold for a high enough price to cover the investment.

Vanity, Information, Misinformation, and Disinformation Domains

Another type of spike pattern we see is people registering vanity domains to create information, misinformation, or disinformation web pages, or possibly even just "for the lulz." These spikes also correspond to notable news events, but what sets them apart from speculation spikes is there is no apparent future value in these domains (at least as reflected in registration or hosting patterns). They were registered specifically for some purpose related to the news event.

One of the more notable examples of this is the "standby" spike. On Sep 29, 2020 Donald Trump used the phrase 'stand back and stand by' during the first presidential debate, which kicked off a small flurry of domains registered using the term "standby".



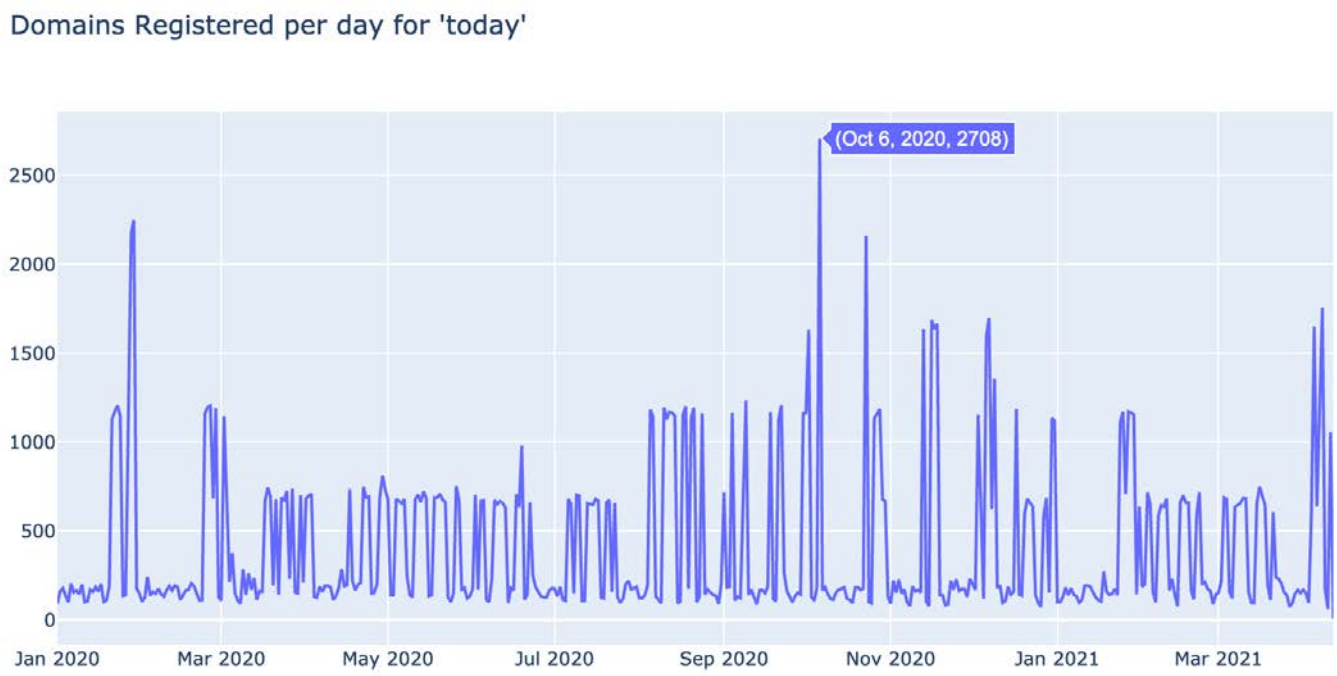
What was interesting about this was that a high number of these domains were just redirects to joebiden[.]com. There was no potential future monetary gain from registering these domains, it seems to have been done for the sake of trolling the other party.

How do we differentiate, in the data, between speculation spikes and political or "lulz" ones? There is no absolute rule, but a reasonable inference can be made based on what kind of hosting is associated with the domains. If the majority of domains in the spike have unique HTML content, return an error or blank web page, or forward to another domain, we infer that the spike was created for non-speculation purposes. We do note that this part of our investigation was carried out via sampling, not programmatic analysis at full scale.

Dictionary Based DGA Spikes

The third type of domain spike that we see quite often are domains that were almost certainly registered according to a dictionary based domain generation algorithm (DDGA). As a matter of fact, some of the original word-frequency analysis that led to this study was born of an effort to find DGAs. DGAs are algorithms that automatically generate a large number of domain names and sometimes, with associated bots, even register them in bulk. DDGAs use a set of hard coded words which are combined in different orders to generate a list of domain names to register.

It is not uncommon for DDGAs to be responsible for several thousand domains in one day, which means the set of words used will often look a lot like spikes. Many DDGA spikes are easy to spot. For example, here is the domain registration timeline for the word "today".



There are several indicators that point to the fact that almost almost every spike is DDGA driven. First, if you look between the spikes, the number of domains registered per day baselines at around 200 per day. If you look at the peak of each spike, almost every spike is approximately 200 domains plus some multiple of 500 domains. Such exact patterns like this do not happen in the wild. These are clear examples of a domain registration automation tool generating 500, 1000, sometimes 2000 domains in one day.

The second way to tell that a spike is a DDGA spike is to quickly scan the domain names in that spike. Here is a sample of domains pulled from the large October 6, 2020 spike.

```
marvelous-advice-letter-to-detect-today[.]info
marvelous-advice-letter-to-gather-today[.]info
marvelous-advice-letter-to-glance-today[.]info
marvelous-advice-letter-to-notice-today[.]info
marvelous-advice-letter-to-peruse-today[.]info
marvelous-advice-letter-to-regard-today[.]info
marvelous-adviceletertocomprehendtoday[.]info
marvelous-announcement-to-examine-today[.]info
marvelous-announcement-to-observe-today[.]info
marvelous-announcementlettertoknowtoday[.]info
marvelous-announcementlettertolooktoday[.]info
marvelous-announcementlettertoreadtoday[.]info
marvelous-announcementlettertoscantoday[.]info
marvelous-announcementlettertoskimtoday[.]info
marvelous-announcementlettertospottoday[.]info
marvelous-announcementlettertoviewtoday[.]info
marvelous-announcementtocomprehendtoday[.]info
```

When sorting the domains by length first, then alphabetically, the DDGA patterns usually group together very coherently and it is then easy to pick out the set of terms that make up the set of terms used by the algorithm.

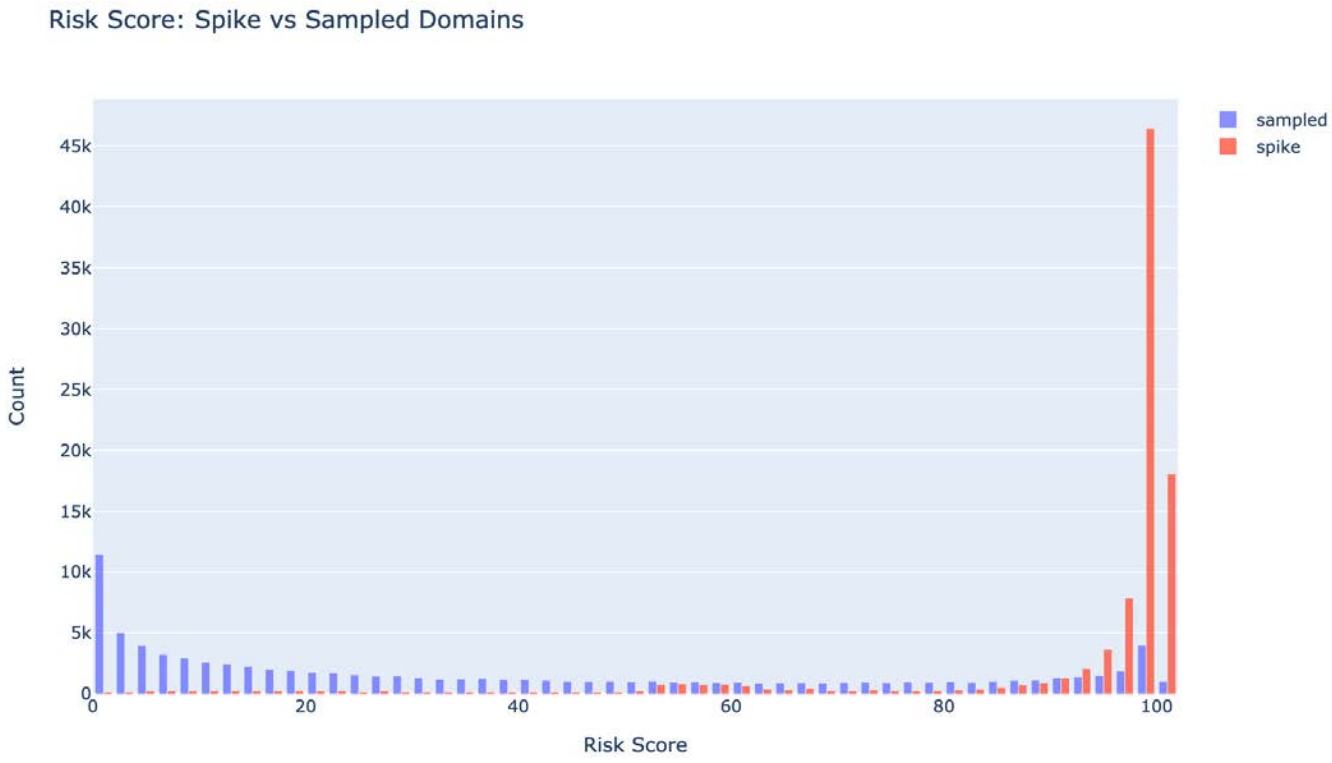
Most importantly, none of these terms correspond to any news events for the day they are registered. These are just random “word salad” domains; traditionally, DGA domains have been strongly associated with malware C2 or with large spam campaigns.

Are Spike Domains Inherently Risky?

Whenever we study large-scale domain registration patterns, we seek to understand whether they reflect a higher risk than the Internet at large. To answer this, we used the following methodology to compare the risk profile of the "today" DDGA domains against a random sampling of domains from the broader population:

- Identify days when spikes occur
- Select all domains registered on those days with the term associated with the spike
- Randomly sample equal number of domains registered on those days that do not belong to the domain spike
- Look up Risk Score for each domain from April 1, 2021
- Plot histogram of Risk Score for the spike domains vs the sampled domains

The graph below plots the risk score distribution for the domains from the "today" DDGA spike against a random sample of domains. The score of zero represents domains that have been proven not malicious, while the score of 100 represents domains that have appeared on industry-approved block lists. The scores in between are the output of the DomainTools Risk Score algorithms. The penultimate red bar indicates that thousands of domains produced by this DDGA are predicted by the Risk Score to be highly likely to be malicious, while the final bar shows that over 15,000 of the domains from the DDGA have already been confirmed bad. It is probable that as time goes by, additional domains that are predicted bad as of this writing, will go on to be blocked, and their scores will go to 100.



We did not observe meaningful differences in risk profiles between the non-DDGA spike domains and the background risk profile for the Internet at large. This was unsurprising for speculation spikes, since such domains typically are not used in any way—they are merely obtained and held. We likewise did not expect to see major risk, as predicted by the DomainTools Risk Score, for mis/disinformation and vanity domains. While disinformation is certainly a dangerous trend on the Internet, it is not currently part of the scoring algorithms for our Risk Score.

We were slightly surprised that other DDGA domains did not show a higher risk profile, since DGA domains are so often associated with malware. Our conjecture is that some of the DDGA domains we studied were either not ever activated (in which case there were no infrastructure components for the Risk Score to process), or were used for relatively benign purposes such as low-value, high-volume mass-marketing campaigns where the "disposable" domains were merely a target for clicks but did not carry any specific risk (other than possible annoyance).

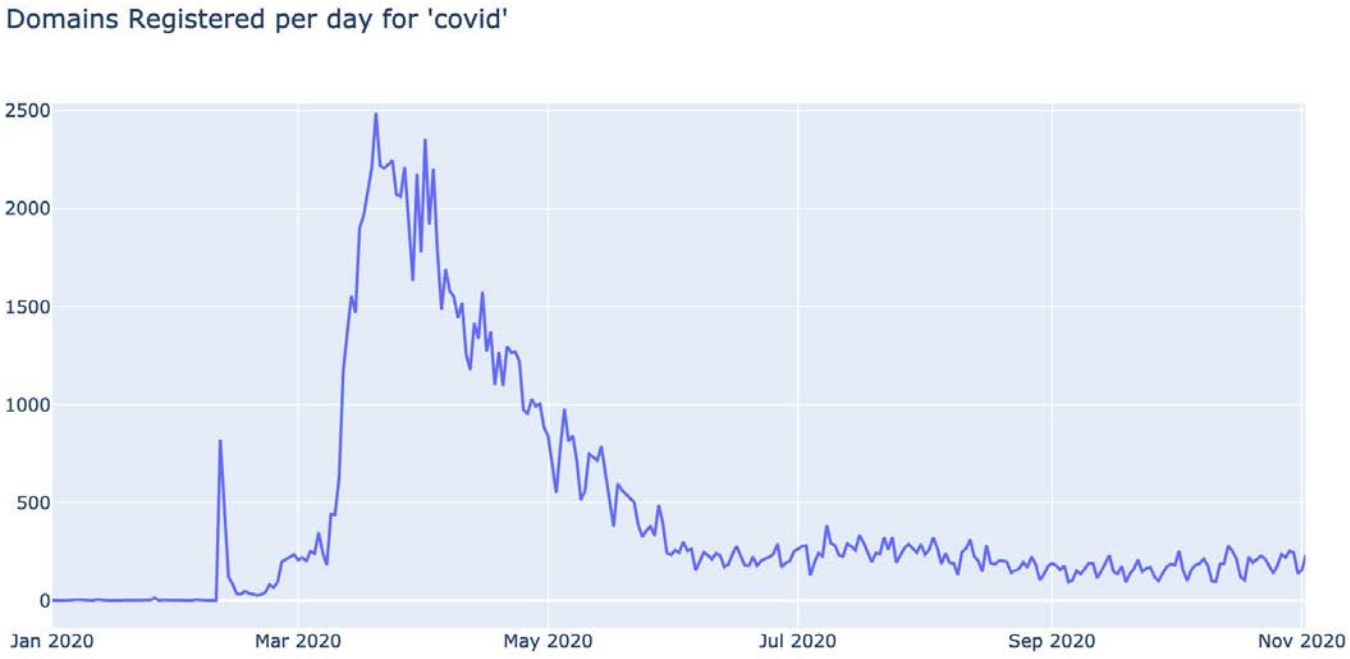
Notable 2020 Domain Blooms

If our assumptions about the causes of domain blooms are correct, it stands to reason that we would observe blooms or bouquets for various major news events of the year. We examined two of those—COVID and Black Lives Matter—to see how they might be reflected in registration patterns.

COVID

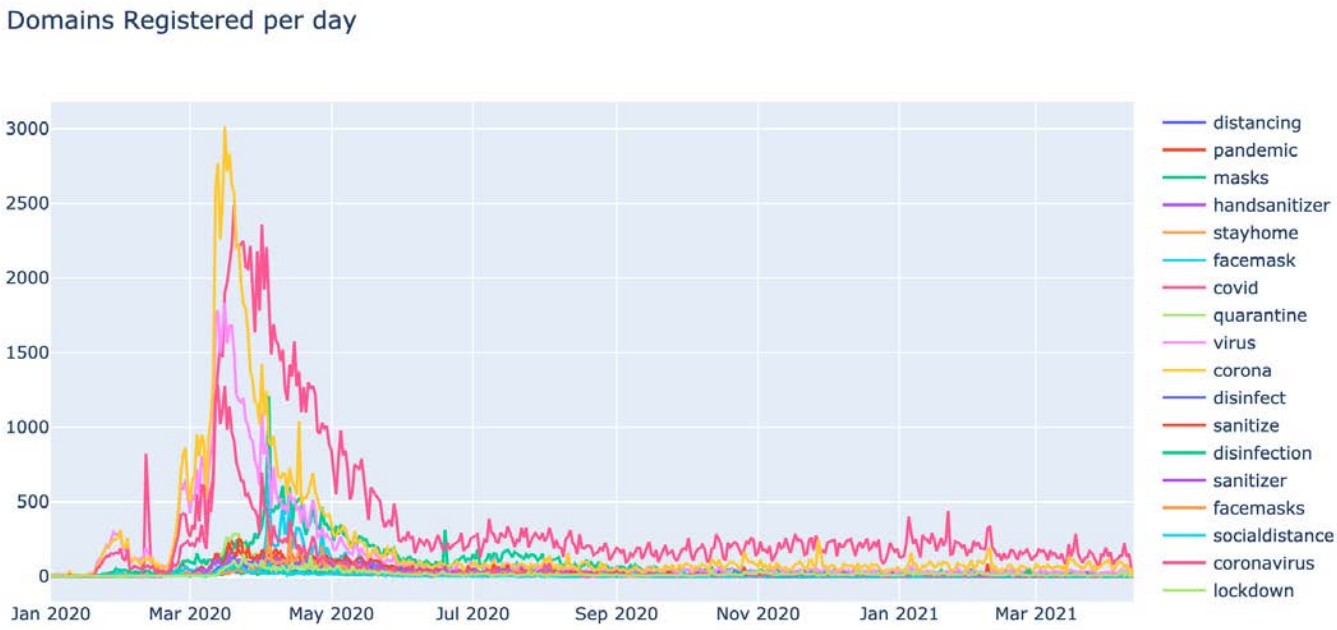
The quintessential example of a domain bloom was covid (see below). Notice that before Feb 11, 2020, the term "covid" was hardly used at all. Then on Feb 11 there was a huge spike in the number of domains using "covid". This domain event corresponds to the World Health Organization (WHO) officially naming the disease "COVID-19" the day before.

After a couple days, the number of "covid" domains per day drops down to about 50 per day. At this point, the "COVID" event looks for all intents and purposes like a domain spike. Then in late February the number of "COVID" domains starts to increase again, but this time they keep increasing until March 12th, at which point they skyrocket to an unprecedented 2,500 domains in one day; this turns out to be the day after the WHO declared COVID-19 a pandemic. For a period of time in March-April 2020, the daily number of "COVID" domains exceeds 2,000.



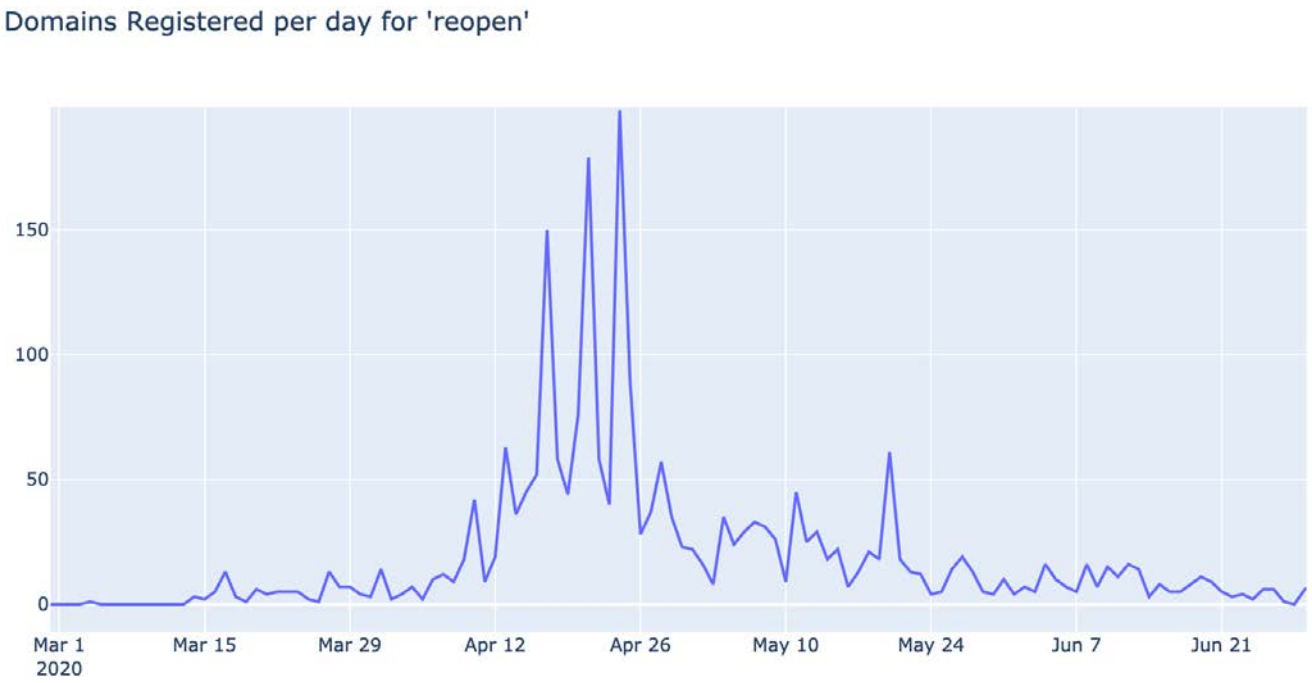
Over time the number of "COVID" domains registered per day gradually drops, settles at a new baseline of around 200 domains per day, and has held steady there ever since.

We noted similar blooms to the above with a series of other words as well, all related to the pandemic or responses to it ("handsanitizer," "facemask," "socialdistance," etc. The scaling of the graph makes it evident that many of the words formed fairly small blooms, but that the patterns were unmistakable even at those low volumes of domains. When we pull the top 18 of these words (in terms of semantic relationship) together into one graph we can see the entirety of the COVID-19 bouquet for 2020.



One interesting COVID-related bloom was for domains containing "reopen." In mid April of 2020, there was what initially seemed to be a grassroots movement to protest the COVID-19 lockdowns and "reopen America".

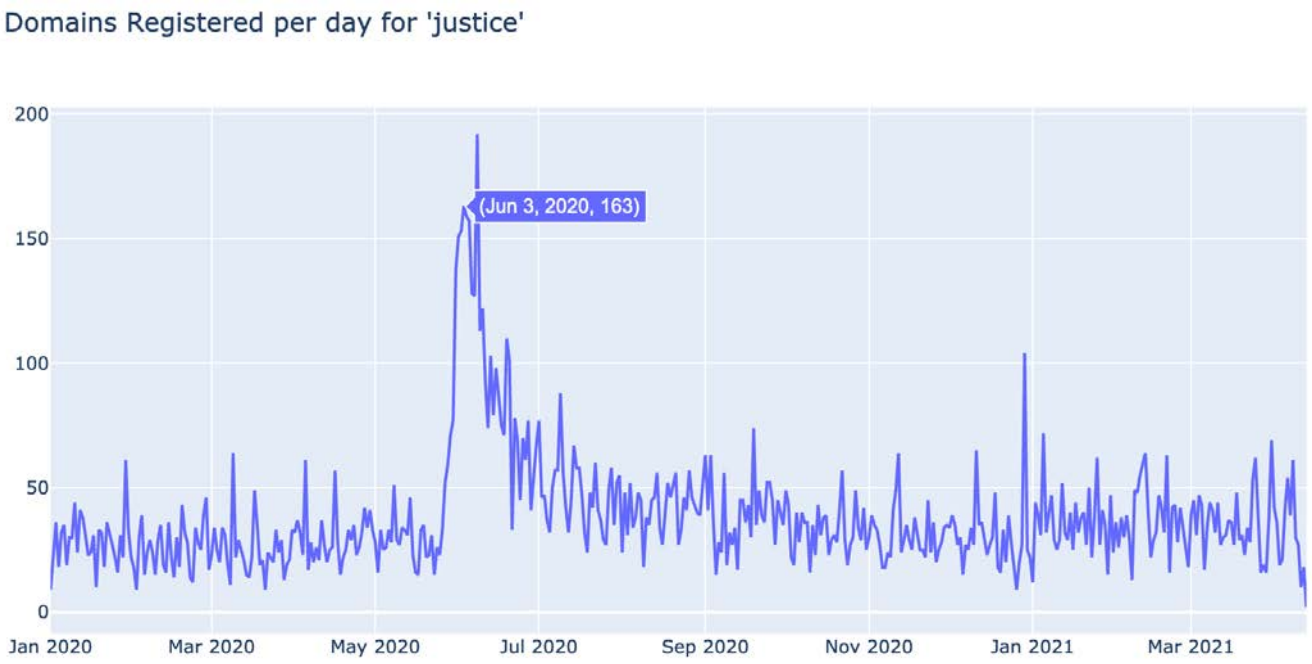
Taking a closer look at the "reopen" pattern, it looks like a legitimate, yet small bloom. It has both volume and duration, but its three main spikes are disproportionately higher than its day to day base number of registrations. This is more of a spike characteristic than a bloom one. Also, the spikes are 3-4 days apart from each other, yet proportionally increasing with each one. This is suggestive of a manufactured pattern as opposed to an "in the wild," grassroots campaign.



DomainTools Security Researcher Chad Anderson did an in depth investigation into these domains and showed they were really a coordinated astroturfing campaign organized by the Dorr brothers.

Black Lives Matter

In late September 2020, the DomainTools Research Team was looking for disinformation domains about the late US Supreme Court Justice Ruth Bader Ginsburg. While looking for domain blooms for these terms, we noticed only one bloom, for the word "justice." However, this bloom started in late May, four months before Ruth Bader Ginsburg passed away.



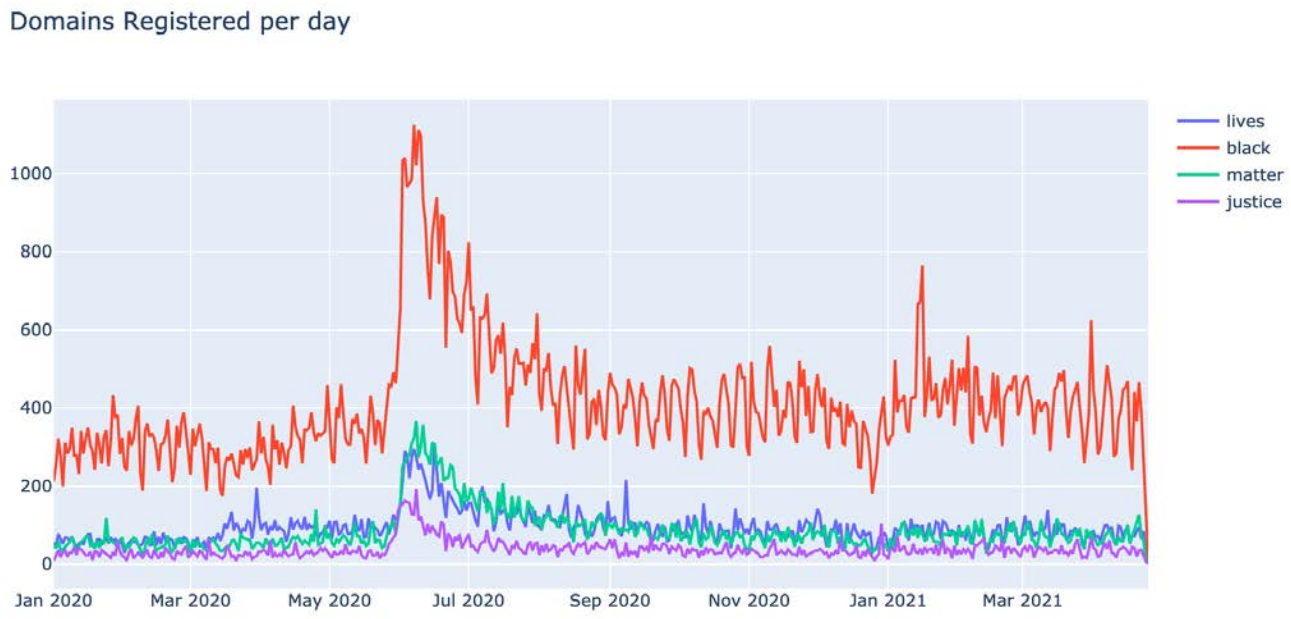
Digging into this bloom deeper, we noticed that it starts to slowly grow on May 27th and then really takes off on May 30th. Inspecting the domains registered during this range it became very clear that we had found a bloom for the Black Lives Matter movement.

justiceforgeorgfloyd[.]com	justiceforbidfloyd[.]com
justicenowgeorgefloyd[.]com	thefightforjustice[.]org
justiceforgeorge-floyd[.]com	breatheforjustice[.]com
justice4georgefloyd[.]org	takeaknee4justice[.]com
takeakneeforjustice[.]com	talksocialjustice[.]com
georgefloydjustice[.]com	

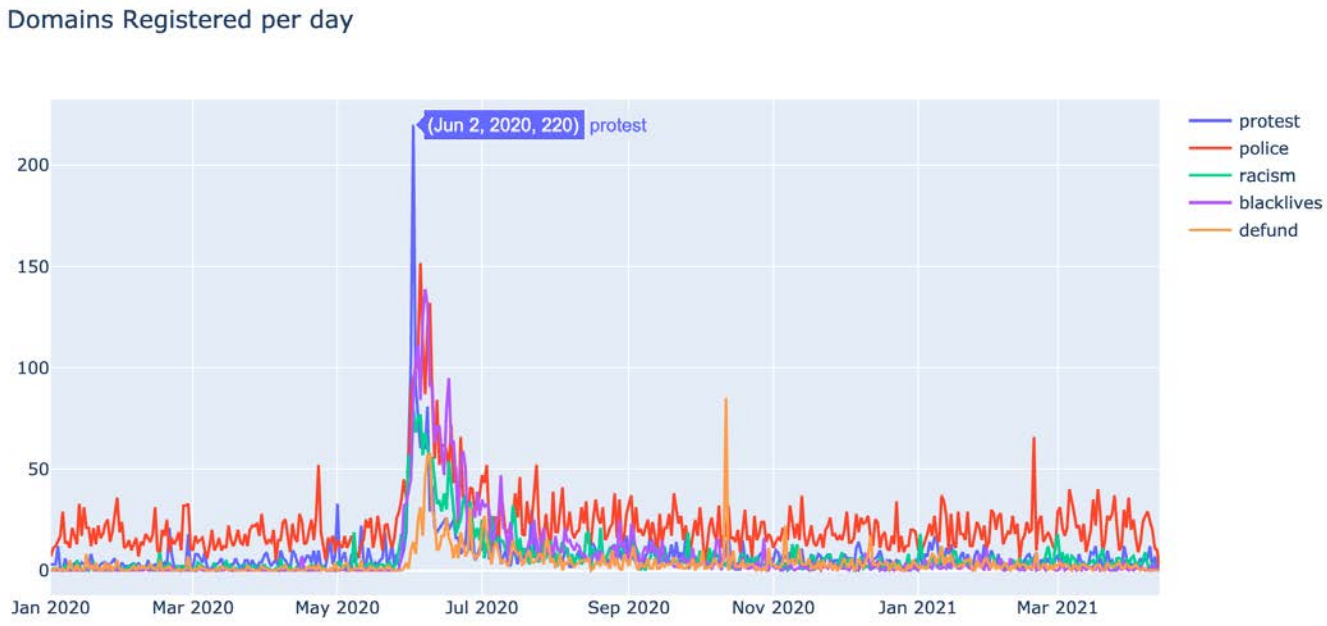
This bloom holds at its peak for about 10 days, before returning down to its new baseline level by early July.

Since George Floyd's name appeared in several of these domains, we also searched for blooms based on the names of other African American victims of police brutality in 2020, but we found a low number of domains, and no identifiable bloom patterns.

When we look up the blooms for the words Black Lives Matter, they correlate perfectly with the "justice" bloom, though they have a much higher magnitude and a longer duration.



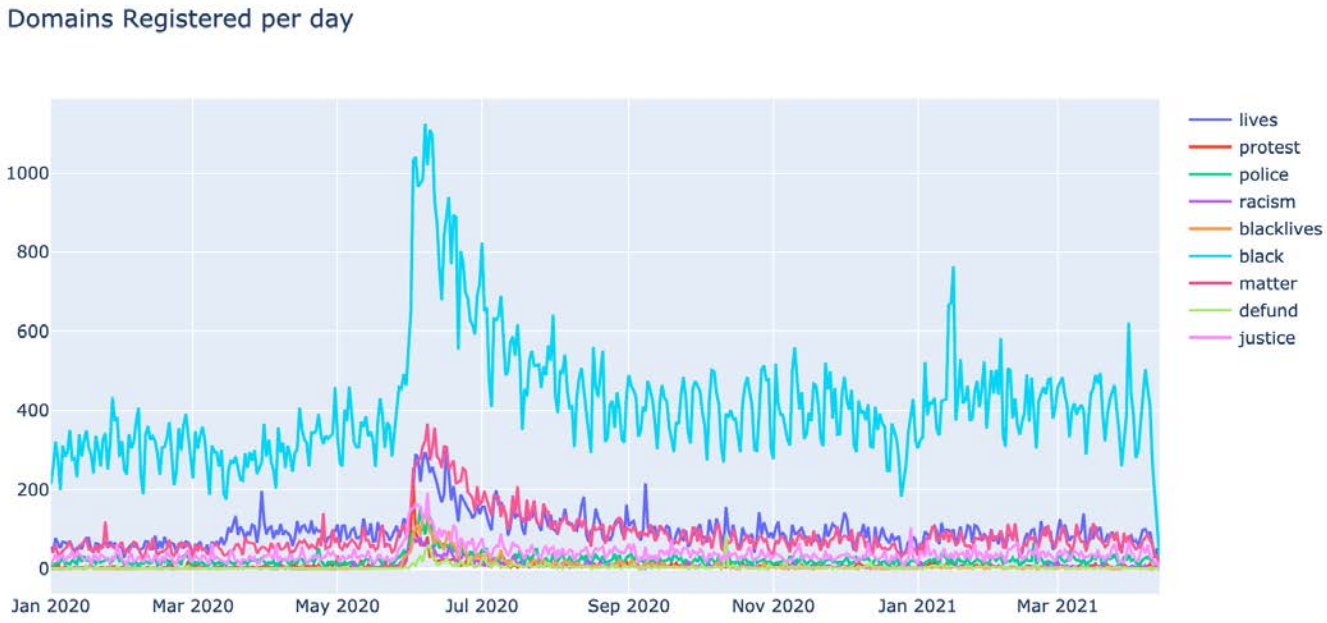
The next major stage of this bloom happened on June 1st when two independent autopsies ruled George Floyd's death a homicide. The next day, as protests raged across the country, we can see the reaction to this news in domain registrations which lasts pretty much through the end of June.



There is an outlier spike on October 12th for the word "defund" that doesn't correlate with the main bloom pattern. Digging into these domains we can see someone just registered a bunch of defund domains targeting random large corporations.

```
defundequifax[.]com
defundequifax[.]net
defundfoxconn[.]com
defundfoxconn[.]net
defundwalmart[.]com
defundwalmart[.]net
defundspirit[.]com
defundspirit[.]net
defundsprint[.]com
defundsprint[.]net
defundunited[.]com
defundunited[.]net
defundapple[.]com
defundapple[.]net
defundexxon[.]com
defundexxon[.]net
defundintel[.]com
defundintel[.]net
defunddell[.]com
defunddell[.]net
```

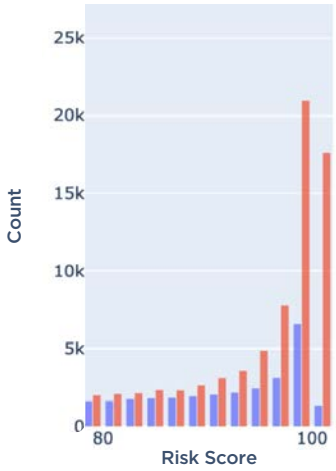
When we pull all these words together into one graph we can see the entirety of the Black Lives Matter bouquet for 2020.



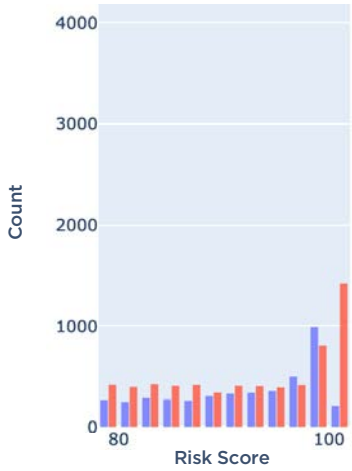
Are Bloom Domains Inherently Risky?

As discussed in the section on spikes, we sought to understand whether domains in blooms or bouquets represent higher risk than the background risk levels of the Internet as a whole. The methodology we used was very similar, except that the time periods for sampling represented the periods of the blooms studied. The graphs below indicate the risk distributions for several domain blooms.

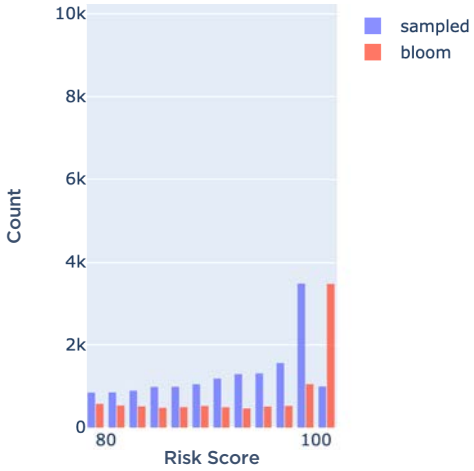
COVID
Bouquet Terms:
pandemic, covid, virus, corona, coronavirus



COVID
Bouquet Terms:
masks, facemask, facemasks



Black Lives Matter
Bouquet Terms:
black, lives, matter, justice



Conclusion

Domains are not registered randomly. Even DGAs, which generate random domain names, don't register domains themselves. Humans register domains for various reasons, and sometimes do so in numbers large enough to generate observable patterns among the ~380 million domains that exist as of this writing. The patterns of blooms and spikes that the DomainTools Research Team identified in 2020 show us that blooms, and some spikes, are strongly correlated with major societal events. We can confidently predict that as other significant events unfold, there will be corresponding new blooms and spikes. By looking at the magnitudes of them, we can make inferences about which events have a relatively larger or smaller value (whether monetary, political, etc). Spikes, unlike blooms, have some specific motivations, including spam or malware campaigns that use DGA domains, speculation based on perceived monetary value of "hot properties," and short-lived political or informational objectives.

From a security perspective, most of the blooms and bouquets of blooms that we studied for this report do not seem to carry a meaningfully different risk profile from the Internet as a whole, **but the COVID-themed bouquet tells us that there will be some cases where many high-risk domains that capitalize on public attention will be created**. We believe the key indicator that a bloom will have a high risk profile is the capacity for threat actors to be able to capitalize on the opportunities presented within the bloom (whether for phishing campaigns, malicious apps, or other purposes). COVID-19 presented many such opportunities, but there really wasn't any way for threat actors to monetize the Black Lives Matter movement.

DDGA spikes have a mixed record, but many of them appear to be malware-related, such as the domains from the "today" DDGA. We intend to continue our research to better understand bloom and spike characteristics so we can classify them as things threat actors will capitalize on (or not). As far as the individual domains within spikes, blooms, or bouquets are concerned, predictive scoring of all new domains continues to be informative in helping defenders and researchers sort out which domains to treat with caution.

Recommendations for Practitioners

The pattern of a bloom or spike is not one that would be observable in the network or endpoint traffic that practitioners typically see during hunting, incident response, or other SOC or intelligence activities. However, it does pay to be mindful of the ways in which malicious actors capitalize on events in the news to attempt to capture attention and, in some cases, gain illicit resources. Thus, we recommend:

- Keep an eye out for event-themed domains in the days or weeks after a significant social, political, economic, or natural event. Domains registered for speculation, for example, will not turn up in logs or SIEM events, but those intended for malicious activities could.
- As with any domains, use characterizing tools such as domain risk scoring to improve situational awareness around event-related domains that do show up.
- Take note of unusually long domain names, as we have seen many such constructions in the DDGA spikes we describe in this report. Some organizations may wish to write some detection rules to pull out domains of unusual length or containing multiple hyphens.

We hope this report proves useful to researchers, security analysts, and other constituencies with an interest in large-scale patterns of domain creation and usage. And as a reminder, for those with an interest in a deeper look at the domain bloom algorithm, we again recommend this blog by John "Turbo" Conwell.