Analyzing the Use of Quick Response Codes in the Wild

Adam Lerner
University of Washington
lerner@cs.washington.edu

Alisha Saxena
Massachusetts Institute of Technology
alishais@mit.edu

Kirk Ouimet
Scan, Inc.
kirk@kirkouimet.com

Ben Turley
Scan, Inc.
ben@scan.me

Anthony Vance
Brigham Young University
anthony@vance.name

Tadayoshi Kohno
University of Washington
yoshi@cs.washington.edu

Franziska Roesner
University of Washington
franzi@cs.washington.edu

ABSTRACT

One- and two-dimensional barcodes, including Quick Response (QR) codes, have become a convenient way to communicate small amounts of information from physical objects to mobile devices. While there is much discussion, awareness, and proposed use of such barcodes, both in academia and in industry, to our knowledge there has not been a systematic and in-depth analysis of the actual ecosystem surrounding these codes. To fill this gap, we analyze a log of all scans performed by users of a popular QR and barcode scanning app available for Android, iPhone, and Windows Phone. Our dataset includes over 87 million scans performed over a 10-month period from May 2013 to March 2014. We examine general use patterns of QR and barcodes in the wild and identify common and uncommon uses and misuses. We see the presence of both conventional (e.g., web) and emerging (e.g., Bitcoin) uses of QR codes, and develop an informed understanding of the types of QR codes being created and how users interact with QR and barcodes in the wild.

Categories and Subject Descriptors

[**Human-centered computing**]: Empirical studies in ubiquitous and mobile computing

Keywords

QR codes; barcodes; empirical studies; mobile computing; ubiquitous computing; mobile malware

1. INTRODUCTION

With the growing prevalence of smartphones and other mobile devices, Quick Response (QR) codes have become a convenient way to quickly communicate a small amount of information, such as a URL, to a user's device. Figure 1 shows a sample QR code; Figures 2 and 3 show examples in real-world contexts. To read these codes, users typically install third-party QR code scanning applications onto their mobile devices. The number and popularity of such applications speaks to the popularity of QR codes themselves. For example, the most popular QR and barcode scanning application for Android boasts over 100 million downloads (as of November 2014, according to `http://appbrain.com`).

Figure 1: Sample QR code.

QR codes are among the most prevalent technologies *bridging the physical and digital worlds*, raising unique opportunities and challenges. QR codes are often associated with physical objects. When a user scans a QR code with a mobile device, that mobile device may perform some follow-on digital action, such as visiting a website (Figure 2) or pairing accounts (Figure 3). Anecdotally, QR codes are used for marketing purposes or to provide pointers to additional information about a physical location or object (e.g., in a museum). QR codes are also used for a growing number of security-sensitive operations, such as authentication, device pairing, and connecting to password-protected WiFi networks. The prevalence and utility of QR codes is likely to increase with the growing adoption of wearable devices such as Google Glass, where text or other traditional forms of input may be cumbersome. For example, Google Glass already utilizes QR codes to connect to password-protected WiFi networks [11]. The research community has also turned to QR codes as a mechanism for linking physical spaces with digital information or computation (e.g., [1, 2, 8, 9, 13, 15, 18, 21, 27–30] — see Section 2 for a discussion).

Though QR and barcode scanning applications and anecdotes about their uses abound, to our knowledge there has been no systematic, in-depth study of their use in the wild. Designers of QR code-based systems are thus forced to rely on speculation, or their own measurements, of the QR code

MobiSys'15, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3494-5/15/05.
http://dx.doi.org/10.1145/2742647.2742650.

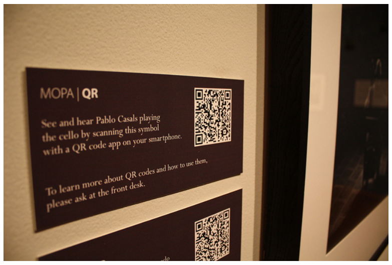
MOPA | QR
See and hear Pablo Casals playing the cello by scanning this symbol with a QR code app on your smartphone.
To learn more about QR codes and how to use them, please ask at the front desk.

Figure 2: A QR code in a museum, encoding the URL for a video related to the exhibit. Image source: `https://www.flickr.com/photos/balboaparkonline/`

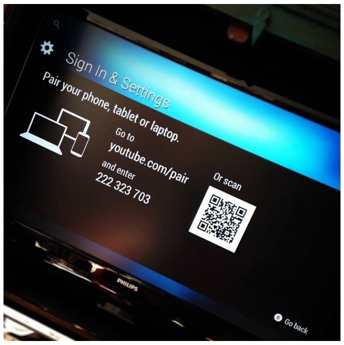

Figure 3: A QR code used to pair a user's YouTube account with the YouTube app on their Xbox. Image source: `http://instagram.com/p/icywcsKZbo/`

ecosystem. To fill this gap, we study the use of QR codes in the wild. We leverage a unique and powerful data set: approximately 87 million scans of QR and barcodes made using Scan (`https://scan.me/`), a popular scanning application with an install base of over 10 million devices.[1] This dataset allows us to examine both the types of QR codes that exist in the wild, and the frequency with which individuals users interact with these codes. We find, for example, that approximately 75% of QR and barcode scans in our dataset lead to web URLs, and that the set of popular websites found in QR codes differs significantly from the set of websites popular on the web in general. We also find that 25% of scans represent other use cases, and we investigate these varied applications (e.g., phone numbers, Bitcoin payments, and two-factor authentication). We also observe examples of malicious uses of QR codes, including links to Android applications containing malware. These cases of misuse are rare, but their presence in our dataset suggests that users may encounter them in the wild.

In this paper, we analyze the use and misuse of QR codes, and we develop an informed understanding of the types of codes that are created and that are encountered by users in practice. Our contributions include:

- We conduct the first (to our knowledge) large-scale academic analysis of QR and barcode use in the wild, using a dataset of 87 million scans. We present general trends about the scans, codes, and devices present in our data.
- We investigate specific use cases for QR codes, including a deep dive into the content of popular codes in our dataset, a comparison between frequently and infrequently scanned codes, and an exploration of the varied use cases present in our data (e.g., cryptographic currencies, device pairing, one-time passwords). We also investigate several potential vectors for malicious QR codes, including malware or phishing URLs and links to malicious Android applications.
- From these findings, we distill a set of lessons and recommendations for QR and barcode scanning application designers as well as for future systems that rely on QR codes or similar techniques to communicate between objects.

We now provide additional background and discuss related work in Section 2 before describing our dataset in more detail in Section 3. We then present general analyses of the dataset (Section 4), and then an analysis of different use cases that manifest in the dataset (Section 5). We discuss the implications of our findings and avenues for future work in Section 6 and then conclude in Section 7.

2. BACKGROUND AND RELATED WORK

One- and two-dimensional barcodes have become popular as a convenient way to quickly communicate a small amount of information, such as a URL, to a user's device. Figures 2 and 3 show examples of Quick Response (QR) codes, a common type of two-dimensional barcode. The prevalence of QR codes has risen alongside the popularity of smartphones [20], with one study showing that European usage of QR codes doubled between 2011 and 2012 [6]. To read these codes, users can install on their devices a variety of third-party QR and/or barcode reading applications, such as the popular ZXing Barcode Scanner, which boasts over 100 million downloads on Android, or Scan (`https://scan.me/`), which has been downloaded over 10 million times for Android.[2]

Research applications. QR and barcodes have been applied in a variety of research contexts. Several efforts leverage the ability of these codes to bridge the physical and digital worlds [2, 29], such as for indoor navigation [27], grocery bargain hunting [8], accessibility [1], and to aid augmented reality applications [15]. Additionally, many security and privacy related uses of QR and barcodes have been proposed, including for communicating privacy policies [4, 28], device-to-device authentication [21], web authentication [13], encryption or verification of real-world paper content [18, 30], and as tattoos of medical device passwords [9].

Security issues. Other researchers have explored the security challenges raised by QR and barcodes, including attacks enabled by ambiguous decoding protocols [7], a study of people's susceptibility to QR code based phishing attacks [32], the potential use of QR codes to spread malware and phishing URLs [35], and other attacks [16]. While these attacks are all technically possible, how prevalent are they in practice? We investigate several types of potentially malicious QR codes in this work, including malicious URLs.

[1] These scans were logged in accordance with Scan's terms of service and privacy policy, and our use of the dataset was approved by our institutions' IRB boards.

[2] Download numbers for Android, according to `http://appbrain.com` in November 2014. The total number of downloads for all platforms is higher.

Table 1: The schema of our dataset. Device UUIDs are random and each corresponds to a single device which has installed the Scan app.

Column	*Example*
Barcode Type	e.g., QR, UPC, etc.
Contents	URI or other data
Location	Lat/lon coordinates
City	City
Region	e.g., state, province, etc.
Country	Country
Platform/Version	Mobile OS and version
Device Type	Phone make and model
Device ID	UUID
Timestamp	Date and time
Scan Source	Camera/History/Gallery

Commercial use. Commercially, QR codes are frequently used for marketing purposes [5, 14], but they have also been applied in various security-sensitive contexts, such as authentication or device pairing. For example, Google experimented with QR codes for passwordless login [25], and Google Glass uses QR codes to connect to password-protected WiFi networks [11]. Indeed, the prevalence and utility of QR codes is likely to increase with the growing adoption of wearable, camera-enabled devices like Google Glass.

Missing knowledge about real-world use. Though QR and barcodes have been frequently applied in a diversity of research and commercial applications, to the best of our knowledge no large-scale academic study has been conducted of the use of such codes in the wild. Thus, we have little concrete knowledge about the prevalence of the various applications and attacks described above. We also have little concrete knowledge about the behavior of real users who may encounter QR codes. We aim to close that gap in this paper, leveraging our unique dataset of 87 million scans from a popular QR and barcode scanning application. In the next section, we detail our dataset.

3. THE DATASET

Origin and Scope. We obtained a log of scans performed by users of Scan (`https://scan.me`), a popular barcode and QR code scanning application for Android, iOS, and Windows Phone. The log includes scans made by real users using the software over a 10 month period from May 2013 to March 2014, including 87,647,504 scans of 18,763,779 distinct barcodes by 15,484,921 distinct devices in 241 countries.

Schema. Table 1 describes the full schema of the dataset. It includes the location and time of each scan as well as an anonymized UUID distinguishing devices which have installed the app. We note that devices are not one-to-one with users, since a user may have multiple devices at once or over time, and a single device may have multiple users. Hence, for expository purposes, when we refer to "users" in this paper, we are often referring to devices.

Human Subjects. The dataset was collected for academic research purposes in accordance with Scan's terms of service and privacy policy (e.g., for Android: `http://scan.me/mobile/apps/scan/android/legal/eula`). We also received IRB approval for this study from the IRBs of the University of Washington and Brigham Young University.

When not stored at Scan according to Scan's policies, we store the data in an encrypted form and performed all of our analyses on machines that we physically control. Though the dataset contains some personally identifiable information (e.g., names and phone numbers in QR codes encoding business cards), we report only aggregate or anonymized data in this paper.

Definitions. Throughout the paper we refer to a "code" as a distinct piece of data contained in a barcode or QR code. For example, Alice and Bob might both embed the url `http://example.com` in a QR code. If a user scans Alice's code and another user scans Bob's code, we consider the *code* `http://example.com` to be scanned twice.

A scan is the act of a user scanning a code with the app, corresponding to one row in the dataset, with all the fields described in the schema (Table 1).

4. GENERAL ANALYSES

We begin by presenting an overview of the barcode and QR code scans in our dataset. We analyze the relative popularity of different types of codes, and examine variations over time and geographical region.

4.1 Basic Distributions

Devices. Our dataset includes 15,484,921 distinct devices, each of which corresponds to a user's mobile device such as a phone or tablet. These devices serve as proxies for users of the app, though we note that devices may not correspond directly to users (e.g., a user may have multiple devices).

We find that a minority of devices account for the majority of scans in our dataset. Specifically, the most prolific 10% of devices performed approximately half of the scans, while just over 30% (4,667,012) performed only one scan over the 10 months covered by our study. Since the dataset is naturally truncated, users who installed the app near the end of our study period will appear to have very few scans. Thus the proportion of infrequent users is slightly exaggerated.

Figure 4 shows the distribution of scans across devices. That figure presents the data over all scans, as well as for only QR code scans and only non-QR code (product barcode) scans. We note that the distribution remains very similar in each of these cases. We speculate that this suggests that the tendency to use a mobile device to scan barcodes is influenced not only by the location and context of barcodes in the environment but also significantly by the individual person's experience and skills — perhaps based upon their affinity to technology. If true, this has implications for adoption of these types of technologies: it may suggest that the uptake of technologies like mobile barcode scanners may depend more on user familiarity or skill than on the ways barcodes are deployed in the environment.

Our results confirm that one- and two-dimensional barcodes can be an effective mechanism for having a physical device influence a digital device for *some* users. For example, 10% of users (accounting for over 1.5 million devices) performed more than 43 million scans in our dataset. The existence of these users supports the use of QR codes in emerging technologies and research projects. On the other hand, QR codes do not currently seem to have sufficient

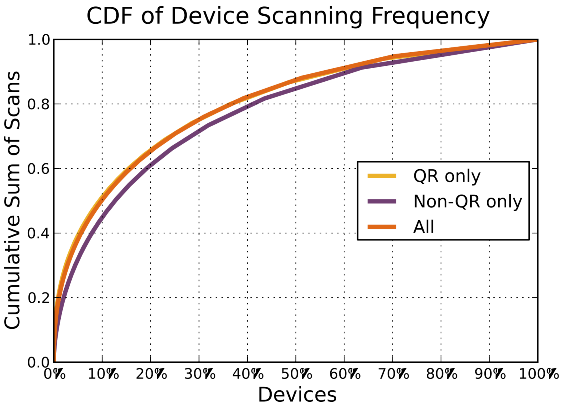

Figure 4: The distribution of scans across devices. Devices had nearly identical distributions of scanning for QR codes and all codes combined — the yellow (QR only) and orange (All) lines overlap almost entirely. The right hand side of the figure shows 30% of devices which contributed only one scan each. Heavy hitters on the left: 10% of devices performed half the total scans in the dataset.

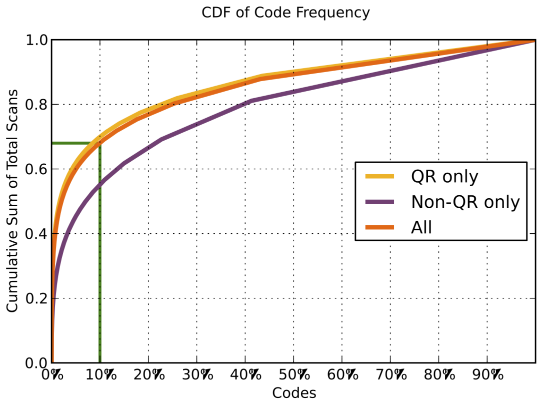

Figure 5: The distribution of code popularity. The top 1% of codes were extremely popular: they made up over 42% of total scans. 10% of codes accounted for over 65% of the total scans that occurred, while in the tail, just over half of codes were only ever scanned once. We see similar trends among all codes, QR codes, and non-QR codes.

appeal for all users (e.g., the 30% of devices with only one scan), thus suggesting that applications that involve a QR-code-based path may not (yet) see much adoption.

Codes. The 18.7 million distinct codes seen in the dataset also followed a typical heavy-hitter vs. long-tail distribution, with a small number of codes scanned many times and many codes scanned only once or a few times. Figure 5 shows the distribution of popularity amongst codes. The most popular code in the dataset had 244,660 scans (0.28% of all scans) while the top 11 individual codes accounted by themselves for 1% of all scans. (We note that the proportion of unpopular codes is artificially inflated by the fact that codes introduced near the end of our time period will necessarily show a low scan count, even if they were subsequently scanned many times.)

Counter to our initial hypotheses, about half of all codes were only ever scanned once, and only a small fraction of codes reach a large number of people. These popular codes speak to a particular set of uses and user experiences of mobile barcode scanning; we explore these popular codes in detail in Section 5.1. For example, we observe that many of the most popular codes are web links to sites of corporations, suggesting that heavy-hitters may be primarily scanned due to their presence in successful marketing efforts.

In addition, we find that less popular codes are more likely to correspond to a different set of applications, including interesting emerging applications such as cryptographic coins, business cards or WiFi pairing. These results thus suggest that QR codes are an attractive tool for designers of emerging mobile technologies. We explore these less popular codes in Sections 5.2 and 5.3.

Code Types and URI Schemes. Scan supports scanning both ordinary barcodes as well as QR codes. QR codes dominate usage of the app: about 87% of all 87 million scans were of QR codes, with the remaining 13% divided between different types of one-dimensional EAN and UPC product barcodes. Table 2 shows this breakdown.

Table 2: Distribution of barcode types appearing in all scans in our dataset. Note that this table counts the appearance of code types in scans, i.e., codes that were scanned more than once are counted once per scan.

Barcode Type	*Count*
QR	76,304,319 (87.06%)
EAN-13	6,554,534 (7.48%)
UPC-A	3,701,269 (4.22%)
EAN-8	687,889 (0.78%)
UPC-E	399,493 (0.46%)
Total	87,647,504 (100%)

While one-dimensional product barcodes encode only numbers, QR codes can contain arbitrary text. When this arbitrary text contains something more than a direct web URL, it is often made more useful by structuring it to contain a URI scheme, such as `tel:` for telephone numbers or `mecard:` for business cards. Table 3 describes the distribution of URI scheme among the scans in our dataset.

In our dataset, we find that the use of QR codes to encode web URLs dominates: as reflected in Table 3, about 86% of QR code scans (about 75% of all scans, including non-QR code scans) contained a web URL. This suggests that quickly connecting a mobile device to a website is by far the most common use case for QR codes. Of these web URLs, we find that only about 10% specified SSL through the `https:` URI scheme.

Though web URLs dominate the QR code scans in our dataset, we nevertheless observe that 14% of QR code scans (about 25% of all scans, including non-QR code scans) contain something other than a web URL. These 10,175,509 scans represent a non-trivial engagement with a variety of other use cases. For example, we did not initially anticipate the prevalence of some URI schemes, such as `wifi:` and `bitcoin:`. We return to discussing such use cases in

Table 3: Distribution of URI schemes appearing in all QR code scans in our dataset. This table shows the number of scans of codes encoding actions in various protocols. Percentages are reported out of the 76 million QR code scans, not the total 87 million scans (that include barcode scans).

URI Scheme	*Note*	*Count*
`http://`		58,488,390 (76%)
`https://`		7,640,420 (10%)
`mecard:/vcard:`	Business cards	1,759,773 (2.3%)
`market:`	Android app store	197,407 (0.25%)
`smsto:`	Send SMS	180,752 (0.23%)
`otpauth:`	Two-factor auth	172,091 (0.22%)
`wifi:`	Connect to Wifi	133,963 (0.17%)
`tel:`	Phone number	123,150 (0.16%)
`bitcoin:`	Crypto currency	39,073 (0.05%)
`itms-services:`	iOS app store	30,663 (0.04%)
`litecoin:`	Crypto currency	11,796 (0.01%)
`dogecoin:`	Crypto currency	317 (<0.01%)
Other		7,526,524 (9.9%)
Total		76,304,319 (100%)

Table 4: Data density of codes of varying frequency of scanning.

	Bytes/code	
Popularity	*Mean*	*Median*
Infrequent (1-5 scans)	58.7	43
Moderate (6-100 scans)	49.9	32
Frequent (>100 scans)	40.9	30

Section 5. A key takeaway from this general analysis, however, is that our dataset provides strong evidence that QR codes are used for many things besides websites.

Data Density. QR codes that encode less information are less visually dense. We investigate the distribution of densities among QR codes in our dataset. We find that on average, more popular codes encode about 18 fewer bytes of information than unpopular codes. The difference is smaller in the median, with a difference of 13 bytes between popular and unpopular codes.

This information suggests that successful, widely scanned QR codes tend to be shorter. We cannot tease apart from this fact whether people who create popular codes tend to make shorter codes (e.g., they tend to include nothing but a URL or explicitly optimized) or whether shorter codes are more likely to be scanned (e.g., a shorter code is processed by scanning applications more quickly, and therefore is more likely to be scanned before the user gives up in frustration).

4.2 Analyses over Time

Usage Trends. Our above analyses suggest that one- and two-dimensional barcodes are effective means to reach some users. We now ask: how do users engage with QR codes over time? For example, once a user installs the app, do they steadily scan codes over time, or does their usage peak initially due to factors like the novelty of the technology?

To investigate this question, we first examined user adoption rate throughout the time period of the data, looking at the first time each device was seen. Adoption rate remained relatively constant throughout the 10 month span — Figure 7 shows the weekly adoption rate by new devices, which hovers below 500,000.

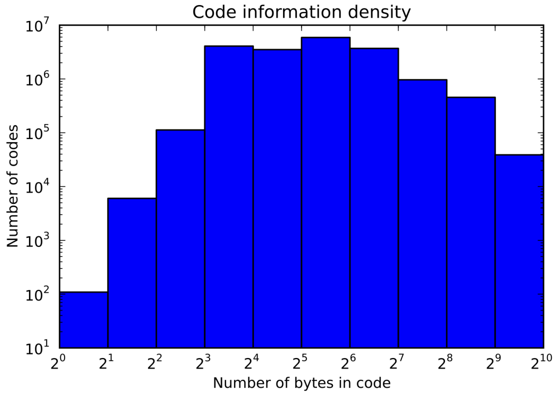

Figure 6: Histogram of data density in codes, on a log scale. Codes of length 8-128 are most common.

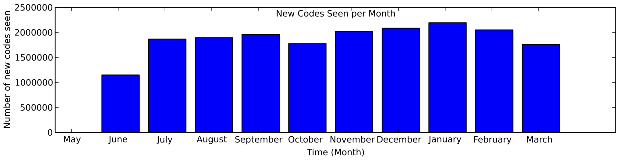

Figure 8: Number of new codes seen per month.

We then examined the number of scans that took place in each week of the studied 10-month time period. The results are also shown in Figure 7. The rate fluctuated from about 1.3 million at the start to about 2.67 million in February of 2014, increasing initially and then leveling off around 1.5 million scans per week.

Comparing the two lines in Figure 7, we observe that although new users appear in the dataset at a regular rate, the number of scans does not increase at the same rate. There are several possible factors that may contribute to this trend besides users whose usage decreases after initial installation and exploration. For example, when a user replaces one device with another, that new device will contribute to the adoption rate but not cause an increased number of scans. Overall, however, this trend suggests that not all users continue to use the app at the same rate after initial installation.

Codes. We also explore the appearance of new codes over time. Figure 8 shows the rate of the appearance of new codes. Each bar shows the number of codes with unique text content, never seen up to that point in time, that were first scanned during that month. The high rate of appearance of new codes suggests that the ecosystem of QR and barcodes present in the physical world is constantly changing.

Note that the rate of new code appearance is quite similar to the rate of new user adoption. This trend suggests an active interest in the QR code ecosystem for both creating new codes and experimenting with the scanning of codes.

4.3 Variations by Geographic Region

Finally, we consider variations by geographic region in our dataset. We use location data reported by devices themselves when they perform a scan.

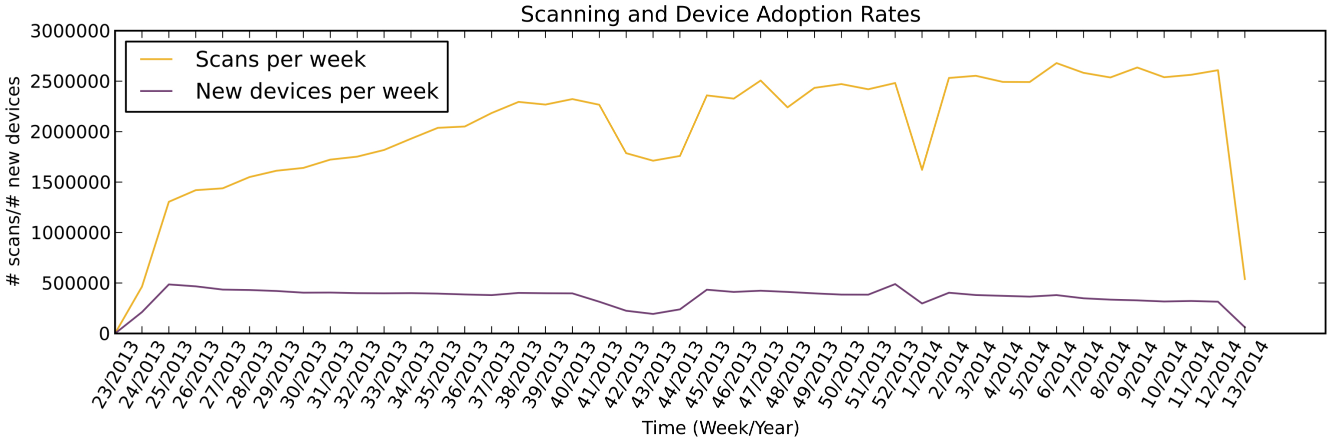

Figure 7: Comparisons between scans per week and new devices appearing in the data per week, which is a proxy for user adoption of the app. These two properties are significantly correlated ($R^2 = 0.295$), suggesting that new users of these types of technologies may be a significant driver of the technology's usage in general. Note that the first and last data points are low because our data for those weeks is incomplete; similarly, the last week of 2013 is low because it is not a full seven days.

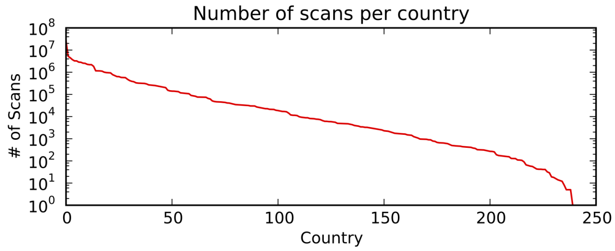

Figure 9: The distribution of scans across the 241 countries seen in the dataset. The United States is at the top with an order of magnitude more scans than its nearest competitor, Germany. Note, however, that the United States has a much larger population than Germany.

Scans. Our data includes occurring in 241 countries. Of those countries, 19 had at least 1,000,000 scans and 60 had at least 100,000. We note that geographical diversity in scanning might be explained by the popularity of different QR and barcode scanning apps rather than the popularity of QR codes themselves. Our vantage with this dataset cannot distinguish such a difference.

Figure 9 shows the distribution of scans across countries worldwide, and Table 5 shows the number of scans in the most popular 32 cities in the dataset. We find that the top-scanning cities are quite geographically diverse, including cities in the United States, Europe, and Asia. These results suggest that QR codes are a global phenomenon, and not something restricted to a particular region of the world.

Codes. Looking across geography, we find that countries vary dramatically in the patterns of their usage of barcodes. Figure 10 shows the ratio of QR code scans to 1D barcode scans in each country as a PDF. Only a small number of countries have more one-dimensional barcode scans than QR scans, but there is notable variance even between those countries which are dominated by QR code scans. For example, QR code scans in one country (Bosnia and Herzegovina) make up about 99% of its 631,812 scans, while scans in Russia (4,482,254 total scans) were split 60%/40% between QR codes and one-dimensional QR codes barcodes.

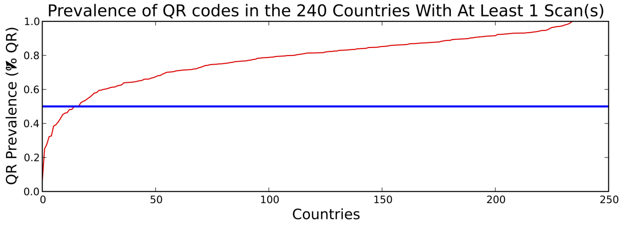

Figure 10: Percentage of scans which were of QR codes (as opposed to 1D barcodes) per country for all countries. The blue line indicates the 50% QR mark — countries below the line performed more 1D barcode scans than QR code scans, whereas countries above the line performed predominantly QR code scans. 6 out of 15 of these 1D dominated countries had fewer than 1000 total scans in the database, and none of the 15 had more than 51,000 scans.

Figure 11 shows the same data limited to the countries in which at least 100,000 scans were performed over the studied period. None of these 59 most scanning countries scanned more one-dimensional barcodes than QR codes, suggesting that QR code based use cases dominate among frequent users of the application. We emphasize that our geographical findings may not be representative of the entire QR code usage ecosystem because our results are only from a single application, and the popularity of this application compared to other applications may vary by geographic region.

Thus far, we have analyzed our dataset as a whole, considering distributions of the frequency of scans and types of codes, analyses over time, and geographic variations. In the next sections, we dive more deeply into specific popular and

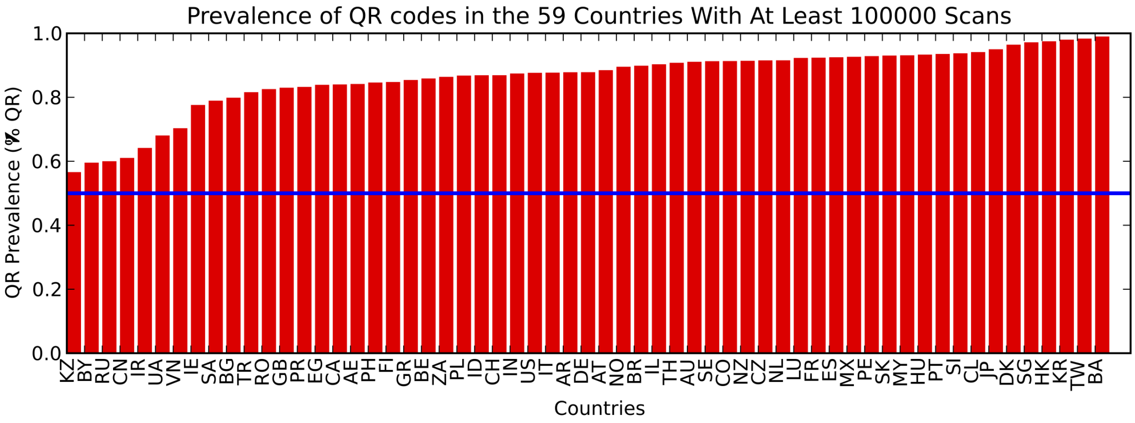

Figure 11: Percentage of scans which were of QR codes (as opposed to 1D barcodes) per country for the most scanning countries. The blue line indicates the y-coordinate for 50% QR codes — all countries with at least 100,000 scans are above the line and performed predominantly QR code scans.

unpopular codes as well as investigate specific use cases of QR codes.

5. USE CASE ANALYSES

Having analyzed the overall properties of the dataset, we now turn to analyzing specific use cases of QR codes and barcodes. Our analysis is three-fold. First, we consider the most popular codes (Section 5.1). These codes reflect the most common uses of QR codes in our dataset, and hence our analysis of them gives insight into a sizeable and important fraction of the one- and two-dimensional barcode ecosystem. However, our dataset contains many codes that are scanned infrequently (half of all codes in the dataset appear only once). Thus, we also conduct an analysis of infrequently scanned codes, which we define as codes that appear 5 or fewer times in our dataset (Section 5.2). Given the diversity of the unpopular codes, our analysis of unpopular codes in Section 5.2 is by necessity different than our analysis of popular codes in Section 5.1. We cannot, for example, simply pick the 100 least popular codes and analyze each of them (indeed, there are 16,792,603 codes scanned 5 or fewer times). Instead, in Section 5.3, we turn to analyzing in more detail specific use cases of QR codes (which span both popular and unpopular codes).

5.1 Popular Codes

5.1.1 Contents of Popular Codes

We first investigate the contents of the 100 most popular (i.e., most frequently scanned) codes. The number of scans of the 100 most popular codes range from 244,660 for the most popular code to 11,925 for the 100th most popular code. The fact that even the most popular codes in the dataset are scanned a relatively modest number of times compared to the total number of scans suggests that users encounter a huge diversity of codes in practice.

We observe that the QR codes real users encounter often contain web links, and that many of the most popular of these links lead to corporate websites. Even among the most popular, however, corporate marketing links do not stand alone, with religious organizations and non-profits like Wikipedia appearing. We also observe a few niche but very popular uses of QR codes, including a code which appears *inside* a video game and a Bitcoin transfer code.

Web Codes. The most frequently scanned codes in the dataset are dominated by links to the web: 95 of the top 100 codes are web links. Most are explicit `http://` links, with a few exceptions: 5 are SSL (`https://`), while 2 are web links that don't specify HTTP or HTTPS explicitly. 40 of the links are to `.com` domains; 8 to `.org`; 3 to `.com.hk` or `.hk`; 4 to `.de`; 5 to `.jp`. At least 15 are shortened links from shortener services and/or QR-code generation services such as `goo.gl`, `bit.ly`, `j.mp`, and `tinyurl.com`, `qrs.ly`, `qr2.it`, and `qrstuff.com`.

The most popular web domains found in the top 100 codes were `jw.org` (Jehovah's Witnesses, 5 of the top 100 codes), `mcd.com`, `mcdonalds.com`, and `happystudio.com` (McDonalds Corporation, including their Happy Studio game, a total of 10/100), and `costco.ms` (3/100). We discuss popular domain in our dataset further below.

Plain Text Codes. Three of the most popular 100 codes are plain text, including the following texts: "**", "tpl_not_defined", and a multi-lingual free text message congratulating the person who scans it for having "successfully identified and scanned a QR code! Great job!" which was included in the video game *Guacamelee! Gold Edition* and was scanned 14,558 times by 8098 different devices. The popularity of QR codes displayed in the game speaks to the viability of mobile system applications that use QR codes as an exchange medium between two devices.

The meanings of "**" and "tpl_not_defined" are unclear to the authors of this study. We hypothesize that the latter may come from the QR code at `http://myopenapps.blogspot.com/2014/04/twilight-war-apk-for-android-free.html`, which appears that it intends to be a QR code that links to the associated application. The resulting QR code may have been created in error and published online before testing.

Other. The only non-QR code in the top 100 codes (#73) is an EAN-8 code for a bottle of a Coca-Cola product, 54491472. The 90th most popular code, with 13,177 scans, is a `bitcoin:`

Table 5: Number of scans in the cities where we observe the most scans.

Count	*Country*	*City*
1458830	TW	Taipei
756964	HK	Central District
436141	RU	Moscow
362383	JP	Tokyo
288478	US	New York
287529	MX	Mexico
235983	CA	Toronto
233458	US	Chicago
223886	US	Austin
223373	US	Houston
220564	US	Minneapolis
194981	SG	Singapore
187634	FR	Paris
180562	GB	London
179355	CA	Montreal
171356	US	San Antonio
170302	US	Las Vegas
164443	DE	Berlin
161584	US	Los Angeles
159301	US	Virginia Beach
153074	KR	Seoul
151308	TW	Nankang
150853	US	Brooklyn
149544	US	Charlotte
143165	DK	Copenhagen
140164	US	Dallas
137590	RU	Saint Petersburg
136527	CH	Full
135315	US	Arlington
134301	US	Washington
133922	DE	Hamburg
130249	US	San Diego

link specifying a Bitcoin transfer to a wallet belonging to `thepiratebay.sx`. We discuss the use of QR codes for Bitcoin and other cryptocurrencies in Section 5.3 below.

5.1.2 Popular Domains

As discussed above, a majority of QR codes contain web URLs. We break these URLs down by domain: Table 6 shows the top 50 domains in codes, ordered by the number of times each domain appeared in a code. Table 6 also shows the Alexa global rank of each domain for comparison.[3]

Counter to our initial expectation, we find little correspondence between domains that are popular on Alexa and domains that are popular in our dataset. While we do see some of the expected popular sites, including `google.com` and `facebook.com`, we also see a large presence of sites that are unpopular on the web in general. Indeed, 15 of the top 50 domains in our dataset do not even appear in the top 1 million sites on Alexa; an additional 7 domains do not appear in the Alexa top 100,000.

Instead of the conventionally popular websites, we observe increased prevalence of URL shorteners (such as `goo.gl` and `bit.ly`), domains that appear to be specific to QR codes

Table 6: The most popular domains found in URLs in codes scanned. These counts are of distinct codes which included one of these domains — multiple scans of a code are not counted here. Note that the count for `google.com` includes subdomains such as `play.google.com` and `docs.google.com`. The right-hand column shows the global Alexa rank for each domain as of November 2014 (values of N/A mean that Alexa does not have data for this domain).

Domain	*Unique Codes*	*Alexa Rank*
goo.gl	2732005	462
youtube.com	2474840	3
google.com	1685677	1
bit.ly	1453335	4406
facebook.com	1436226	2
apple.com	1266273	35
qrs.ly	1063665	2789080
kaywa.me	689613	3124084
delivr.com	566530	243915
premier-kladionica.com	527672	133052
scn.by	507066	N/A
645lotto.net	374325	288448
youtu.be	369834	10515
mcd.com	347196	152098
jw.org	341542	1415
qq.com	316432	10
naver.com	311335	112
bitly.com	309755	388
mta.info	293387	6151
scan.me	277870	140254
nlotto.co.kr	269148	57408
vqr.mx	266588	3349845
tagr.com	256771	13590936
naver.jp	250224	187
towerofsaviors.com	249296	98650
mon-gain.fr	239418	5561369
dropbox.com	237684	85
metrohk.com.hk	187164	81666
bby.us	181854	11907472
tinyurl.com	165680	591
2d-co.de	163990	17873277
augme.com	160901	N/A
safeshare.tv	160412	17952
windowsphone.com	158093	1245
phonegap.com	150780	6809
q-r.to	137553	2758289
happystudio.com	137437	98193
tipico.com	133337	10734
wikipedia.org	126085	6
mcdonalds.de	122256	30767
azon.biz	109331	N/A
j.mp	106171	159932
vk.com	105176	22
opn.to	104969	8613897
notifii.com	104804	2065475
padlet.com	101033	6232
qr2.it	99324	16832364
costco.ms	98246	N/A
paninifootballleague.com	95096	137979
bokgwon.or.kr	93588	N/A

[3] Domain rank data acquired from `http://www.alexa.com` in November 2014. The domain ranks may have differed at the time the codes containing the domains were scanned.

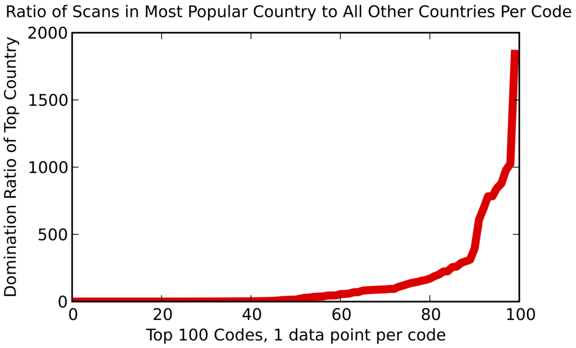

Figure 12: The degree of geographic diversity in scans of the top 100 codes. The y-axis represents the degree to which the most scanning country dominates scans from all other countries, represented as a ratio of scans from that country to scans from all other countries.

(such as `qrs.ly` and `kaywa.me`), domains related to lotteries (such as `645lotto.net` and `nlotto.co.kr`), and more. These results suggest that the QR code-based web ecosystem is different than the traditional browser based web, with some of the major QR code-based web sites not having a proportionately large presence on the browser-based web.

5.1.3 Geography of Popular Codes

The most popular codes in the dataset are quite diverse in terms of geographic diversity. On the one hand, some codes show very high geographic diversity, with the most popular country for one of these popular codes responsible only for about 9% of scans. On the other hand, some codes show almost no geographic diversity. 37 codes are dominated by scans from a single country by a factor of 100: the most popular country for each of these codes has scanned the code 100 times more often than all other countries combined. In the most extreme case, the most popular country for a particular code (Japan) has over 1800 times the number of scans (12900 scans) as all other countries combined (12907 combined scans total).

Figure 12 shows the ratio of top-country scans to all scans for the 100 most popular codes. For a third of the most popular 100 codes, this ratio is under 1, indicating that no single country was responsible for more than half the scans of these codes. The content of these high-diversity codes are quite diverse, including the link to donate bitcoins to The Pirate Bay, the code from the Guacamelee video game, English language Wikipedia, and a link to a personal blog. 11 of these 33 high-diversity codes lead to pages which are primarily offering the download of mobile apps or which include calls to action to download mobile apps. This suggests that codes with high geographical diversity are more likely to be used in at least two particular cases: more unusual/less corporate uses (religious and non-profits), and to encourage installation of mobile applications.

The most popular code in the dataset (244,660 scans) links to a landing page for a mobile game for iOS and Android. Its scans are dominated by scans from Taiwan (243,450 scans in Taiwan), with a ratio of 200 Taiwanese scans to each scan from any other country (443 scans in the next-closest country, Malaysia). Digging deeper, those scans in Taiwan are dominated by scans in Taipai, but the code retains significant geographic diversity of scans within Taiwan.

5.2 Unpopular Codes

Having considered popular codes, we now turn to unpopular codes. In particular, we consider infrequently scanned codes, which we define to be those that appear 5 or fewer times in our dataset. Note that while the vast majority of codes are infrequently scanned (89% of codes), the scans of those codes account for only 31% of all scans in the dataset, since the frequently scanned codes are often scanned hundreds or thousands of times (see Figure 5).

Given the large number of infrequent codes (16,792,603) and their diversity, we cannot simply pick the 100 least frequently scanned codes (as we did for popular codes) and analyze each of them. Rather, we explore now the quantitative differences between frequently scanned, moderately scanned, and infrequently scanned codes.

Specifically, we compare trends of content and usage among codes of differing scanning frequency. For this analysis, we defined "frequently" scanned codes to be those scanned more than 100 times in our dataset, "moderately" scanned codes to be those scanned 6-100 times, and "infrequently" scanned codes to be those scanned 5 or fewer times. This division into bins of 1-5, 6-100 and over 100 scans serves to divide the codes into groups each accounting for approximately one third of all scans in the dataset.

Table 7 lists the fraction of codes for each content type (e.g., web, telephone, etc.) that were infrequently, moderately, or frequently scanned. The important comparisons in this table are not across columns — since most codes are scanned infrequently, a given code is most likely to be an infrequent code. Instead, this table allows us to compare across content types in a column: is one content type more or less likely than another content type to be in that frequency category? For example, Android `market:` QR codes are more likely to be scanned moderately than business cards.

5.3 Specific Use Cases

We now turn to our analysis of specific uses — and *mis*uses — of QR codes present in our dataset. These use cases, which span both popular and unpopular codes, present a snapshot of different actual uses of QR codes in the wild.

We organize this section into subsections corresponding to five themes: web-related use cases, codes including private data, emerging uses, errors, and malicious uses. A single use case may in fact span multiple themes, but we have organized the section this way to highlight these themes.

5.3.1 Specific Web-Related Uses

We dive more deeply into two specific web-related use cases: shortened URLs and links to adult websites.

Shortened URLs. Many of the web URLs scanned by users were shortened by one of several URL shortening services (e.g., `bit.ly` or `goo.gl`). 984,447 of 11,763,834, or about 8% of URLs seen in the dataset were shortened. To find shortened URLs we checked them against a list of about 40 shortening services that we compiled from sources on the Internet and manually checking URLs which looked like they might be shorteners. A small number of URL shortening services dominated the shorteners seen: 88% of shortened URLs were shortened by Google's URL shortener (`goo.gl`) or by Bitly (`bit.ly`, `bitly.com`). `qr.net`, a service which

Table 7: Comparison of QR code contents for three categories of code: frequent, moderate and infrequent codes. We defined infrequent codes to be those with 1-5 scans, moderate codes those with 6-100, and frequent codes those with over 100 scans. The table lists the fraction of codes for each content type that were infrequently, moderately, or frequently scanned. For example, 18.15% of codes leading to Android apps on the market (last row, `market:`) had 6-100 scans (moderate), while a business card is only 1.78% likely to be scanned that many times. The dominance of infrequent codes in general is due to the fact that most codes in the dataset are infrequently scanned — 89% of all codes had 5 or fewer scans (see Figure 5).

Scheme	Total Codes	Frequent %	Moderate %	Infrequent %
Business cards (vcard + mecard)	670359	0.01%	1.78%	98.20%
Dogecoin	184	0.00%	3.26%	96.74%
BBM (Blackberry Message)	71985	0.01%	3.45%	96.54%
Litecoin	525	0.19%	4.38%	95.43%
otpauth (2-factor auth)	79000	0.00%	4.87%	95.13%
Bitcoin	10199	0.16%	4.98%	94.86%
Wifi	55741	0.03%	6.58%	93.40%
iOS apps (`itms-services:`)	10179	0.14%	7.51%	92.36%
Barcode	3647582	0.00%	8.14%	91.86%
Telephone #	38849	0.13%	8.62%	91.26%
HTTPS	1507300	0.43%	9.88%	89.69%
SMS	839	0.12%	10.73%	89.15%
HTTP	12254403	0.53%	11.13%	88.33%
Android Market (`market:`)	18544	1.15%	18.15%	80.70%
All (mean over all contents)	18763779	0.4%	10.2%	89.4%

simultaneously shortens a URL and generates a QR code linking to that shortened URL, appeared as the 7th most popular shortener with 4224 shortened URLs.

We can hypothesize a few purposes for shortening URLs in QR codes. While it may seem odd to shorten URLs in a format which is already only machine readable, we note that URL shorteners are often used for analytics purposes (e.g., to track clicks). Additionally, QR codes with shorter URLs will have less data density, and hence may be easier to scan by some QR code scanners.

Adult Websites. We observed a number of scans of URLs leading to websites which serve adult content. We manually identified 4 domains in the Alexa top 100 whose names clearly indicate that they are adult sites (we may have missed adult sites with less obvious names). We found 7736 scans of 1929 distinct links to these 4 domains.

The human-opaque nature of QR codes makes them a vector for displaying or referencing age- or context-inappropriate material in plain view. We note that for many popular barcode scanners, including the most popular one in the Android Marketplace, the scanner app automatically fetches and displays the title of the specified website. Thus, even if the app doesn't automatically take the user to the site, and even if the site asks visitors for their age (the most popular site in the dataset does not), users will be shown the title of the video or page linked to, which may be inappropriate for them or their context. Further, automatically fetching the website's title will cause the user's device to connect to the site in question without the user's explicit intention.

Our dataset shows that QR codes with links to adult content do exist and are scanned by real users with some regularity. We thus suggest that the designers of QR code scanning applications consider these concerns, perhaps using a list of known adult websites to mask website titles, avoid automatically fetching titles at all, or show additional warnings before redirecting users to such sites.

5.3.2 *Private Data*

We observe a number of use cases that involve encoding private information into a QR code.

Wifi Setup Codes. QR codes can be used to encode the information needed to connect a device to a Wifi network. We found 55,809 unique such codes, which were scanned a total of 134,121 times.

For private Wifi networks, these codes contain the plaintext password used to authenticate with the network, exposing that password to anyone with a QR code reader in range of the QR code. As always-on devices with the capability to read QR codes become more prevalent (e.g., Google Glass), we might expect that the threat of (intentional or unintentional) shoulder surfing to read QR codes containing such sensitive information will increase. Another lesson from these results is that designers of applications that emit QR codes should consider the implications of putting private information in the QR codes — an untrustworthy QR code scanner would be able to extract that information. The makers of QR code scanners must also take precautions to protect the privacy of data contained in scanned QR codes.

Two-Factor Authentication. We found 79,017 distinct codes using the `otpauth://` scheme, which is used to set up two-factor authenticators (e.g., Google Authenticator). This suggests that two-factor authentication is used by a significant number of people. The URIs included codes to set up authenticators for Microsoft-related accounts (25,753), as well as accounts for Facebook (15,977), Gmail (10,108), Dropbox (2360), Zoho (766), WordPress (741), GitHub (524 codes), DigitalOcean (427), and a large number of others, some of which appeared to be malformed. Most of these codes (78,680) used the standard time-based (TOTP) version of the protocol; the remaining well-formed URIs specify the HMAC-based (HOTP) version of the protocol.

Of the `otpauth://` codes we found, the vast majority (about 99.5%) were only ever scanned by a single device, indicating that only one device was set up to authenticate for that account. Note that unlike Wifi setup codes containing passwords (discussed above), viewing another user's two-factor authentication setup code is not as dangerous, as they do not allow the attacker to compromise the user's account without the primary authentication password. However, they would allow someone to compromise the user's second factor by obtaining the secret used to initiate it.

PGP Keys. A QR code containing a public key such as a PGP key is a natural way to convey cryptographic credentials. For example, a person might include their public key as a QR code on their business card to make it easy for acquaintances to import the key and support subsequent communications. Key distribution has long been a topic of interest and study for making cryptographic systems of trust usable and used. Physical artifacts such as QR codes represent an interesting point in the space of these solutions.

The dataset contains 30 codes representing PGP public keys and 4 codes representing private keys. One of these PGP keys is found embedded in a business card format, while the others stand alone, with the entire QR code representing a full PGP public key block. The public keys were scanned only a total of 56 times, representing a tiny percentage of the scans in the data. The 4 private keys were scanned only 7 times altogether. While we cannot conclude anything in particular from these examples, we note that sharing a private key instead of a public key could be a serious breach of cryptographic security, and that code creator error could result in the inclusion of private keys in scenarios like QR code generation. While these inclusions of private keys could be intentional, the danger of human error in QR code creation is illustrated here. As with Wifi and two-factor setup codes, this finding suggests that QR code creators and consumers should exercise care with private data.

5.3.3 *Emerging and Niche Uses*

Bitcoin and Other Crypto Currencies. Our dataset provides us with a glimpse into the use of Bitcoin [24] and other crypto currencies, such as Dogecoin [10] and Litecoin [26]. The `bitcoin:` URI scheme is used for directing a device to make a Bitcoin payment and includes the wallet to pay to and the number of bitcoins to transfer. We found 10,199 codes containing Bitcoin URIs, specifying payments to 8541 distinct Bitcoin wallets. The most popular Bitcoin URI in our data points to `thepiratebay.se`'s Bitcoin address, suggesting that many people made (or considered making) donations to that site. This URI was popular enough to be one of the 100 most scanned codes. Note, however, that the presence of a Bitcoin transaction scan in our dataset does not necessarily mean that the transaction was confirmed by the user and committed to the Bitcoin network.

Bitcoin is significantly more prevalent in our dataset than other cryptocurrencies; we found only 186 Dogecoin URIs and 525 Litecoin URIs.

We also found codes which appear to represent Bitcoin private keys in Wallet Import Format (WIF) [3]. Keys in WIF are 51 characters long and begin with 5 for private keys. We found 2483 codes in the format of a bitcoin private key, according to the above definition, and verified that 1260 of them are well formatted, i.e., that they can be decoded into private keys which could be used to import the corresponding wallet. We did not import any of them.

Boarding Passes and Event E-Tickets. We found electronic tickets, both for transportation as well as for events such as concerts, among the codes scanned in the dataset.

For boarding passes, we found 2416 codes which appear to be in a standardized format for airline boarding passes containing confirmation codes, flight departure and arrival times, airport codes, and names of passengers. These boarding passes were scanned a total of 5396, with 1101 of them scanned only once, 668 scanned twice, and the remaining 647 were scanned more than twice. One boarding pass was scanned 25 times by the same device.

We also found a variety of URLs and eTicket formats for event tickets. We note significant diversity in the formats and strategies used by eTicketing systems. For example, some codes included an `https://` link to a backend ticket processing system which doesn't seem to host public content , while others lead to sites on the public web. These systems seemed not to follow a common standard or format.

There is greater standardization in airline boarding passes than in other eTicketing systems. We attribute this naturally to the need for interoperability between different airlines, security agencies and airports. The contrast between these two systems with similar purposes (admission of a person to a restricted location or event) but differing levels of standardization speaks to the different ways that technologies like QR codes can be used depending on usage context.

Never Gonna Give You Up. A popular joke on the Internet is to link unexpectedly to a video of the Rick Astley song "Never Going to Give You Up" (referred to as "rickrolling") [34]. Because QR codes are not human-readable, they may be an appealing mechanism for delivering this URL to an unsuspecting victim. Indeed, we found 1614 scans of 40 un-shortened codes and 24 shortened codes containing the URL of the video (`https://www.youtube.com/watch?v=dQw4w9WgXcQ`) by devices in 63 countries. The simplest code for a rickroll (with only the URL presented above) was scanned over a thousand times and was likely created by many different people who all chose to create identical codes which were scanned in a diversity of places and at differing times. The most popular country for rickrolling was the United States (with 589 scans), followed by Great Britain and France (with 194 and 154 scans respectively). This suggests that while a few dozen people may have thought to use QR codes in such a joke, each individual creator is unlikely to have spread their code far. While these numbers are small, it suggests that people are using QR codes for innovative, homebrew purposes.

Similar in spirit to rickrolling, we find 993 scans, over 711 different devices and 597 different cities, of the following quote from the movie *Fight Club*: "It could be worse. A female could cut off your Dick while your sleeping and throw it out a moving vehicle." (sic) This QR code, which consisted entirely of text and not a link, also demonstrates that QR codes can be used as a vehicle for communicating directly with humans, rather than first to a mobile device for processing (e.g., rather provide a URI or a PGP key).

TagMeNot. Our dataset gives us a unique opportunity to measure the prevalence of emerging uses of QR codes.

For example, TagMeNot.info is a "pre-emptive, anticipatory, vendor independent, and free opt-out technology for pictures taken in public places" [4]. TagMeNot is an early example of cyber-physical interactions in which aspects of the physical world can be interpreted in digital context. QR codes from TagMeNot indicate that the wearer of the code wishes to opt-out of certain sharing or usage of their likeness or property by the takers of photographs. We found that the code TagMeNot.info was scanned only 7 times by 5 different devices in Mexico, Italy, Great Britain, and the United States, suggesting that such opt-out QR codes have not been widely adopted for privacy in the face of ubiquitous cameras.

5.3.4 Erroneous Uses

Throughout our analysis of our dataset, we observed a variety of malformed QR codes. In this section, we consider one potentially erroneous use in detail.

JavaScript. To our surprise, we observed a number of QR codes containing JavaScript or HTML content. For example, 154 codes, which were scanned a collective 303 times, contained JavaScript code under the URI scheme javascript:. Most of these codes are treated merely as plaintext by most QR code readers (i.e., the readers do not attempt to execute the JavaScript, or even offer the opportunity to do so).

We hypothesize that most of these codes suggest a lack of understanding of how to use these code snippets or of the purpose and usage model of QR codes on the part of QR code creators. For example, one of the JavaScript snippets is code for a browser bookmarklet that creates a QR code [17]. Its presence in our dataset suggests that the creator misunderstood how to use this code, which should be pasted into a browser bookmark rather than into a QR code.

5.3.5 Malicious Uses

Finally, we consider a number of potential malicious uses of QR codes and investigate their prevalence in our dataset.

Premium Telephone Numbers. In the United States and Canada, certain telephone number area codes designate sets of numbers as toll-free or premium numbers [31]. One might hypothesize that QR codes are a vector for tricking people into calling expensive premium numbers. However, we did not find any numbers that we believe to be premium numbers from the US/Canadian system. We did find 667 numbers which appear to be toll free numbers (for example, 1-800 numbers). These numbers were overwhelmingly scanned in the US (91% of scans), suggesting that our guess that these numbers are US toll free numbers is correct.

Special Phone Numbers. Phones often respond with special actions, such as displaying statistics or factory resetting, when special codes are dialed. For example, some Samsung phones display their IMEI number when "*#06#" is typed into the dialer. We find 126 scans of Samsung special codes in our dataset. One such code factory resets certain Samsung devices (*2767*3855#) [22]. Putting such a code into a QR code may be dangerous, since the code must only be displayed in the dialer (i.e., the user must not press the call button) and the device does not ask for confirmation. Indeed, we find 17 scans of this code, suggesting that someone may have attempted to test or actually carry out an attack. Fortunately, none of the scans come from devices of the make and model that treats this code as a factory reset.

Malicious URLs. Prior work [32, 35] has suggested that QR codes are a promising vector for distributing malicious URLs. Intuitively, QR codes seem like a natural conduit for phishing attacks or the distribution of malware (e.g., sending users to drive-by-download sites). QR codes are opaque, machine readable pointers, which might make it harder for a user to check that a link is trustworthy. Some QR code apps (including Scan) can be configured to automatically load the URL in a browser without user confirmation. Additionally, the real-world context surrounding a QR code (e.g., its placement on a marketing poster for a trusted brand) might cause the user to trust the code. Our dataset gives us the unique opportunity to study whether users encounter malicious QR code links in the wild.

We investigate this question using Google's Safe Browsing API[4], which provides classifications of a URL as *malware* and/or *phishing*, or *ok*. We randomly sampled URLs in our dataset and queried them against this API (unshortening any URLs shortened by bit.ly in the process, using bit.ly's API). Note that the malware/phishing status of a website may change over time (e.g., as a malicious site is taken down or a legitimate site is compromised), but the Safe Browsing API does not provide us with historical data, so our results are necessarily limited.

Of the 1,017,955 unique URLs that we tested, we found 209 URLs flagged as *malware* (0.02%) and zero URLs flagged as *phishing*. We were surprised by the latter result, which may suggest that QR codes are presently not a common way to distribute phishing websites. The 209 malicious URLs represent 106 unique domains. Though the fraction of malicious URLs in our dataset is low, the fact that we observe some instances of potentially dangerous URLs suggests that QR code scanning applications should integrate the Safe Browsing API or similar to check URLs before automatically visiting them or allowing the user to visit them.

Note that like QR codes, shortened URLs hide the destination URL and thus might serve as convenient vectors for distributing malware and phishing links. However, similar to our findings, prior work has found that users rarely encounter malicious shortened URLs [19].

Malicious Android Applications. Digging deeper into the many web URLs in the dataset, we find that a non-trivial number point to Android applications (i.e., apk files). Of QR codes containing web URLs, 49,282 (0.07%) contain such links. Whereas the market: URI scheme points to Android applications on the official Google Play app store, apk files referenced by web URLs do not come from the official store. Users who have enabled the setting on their Android device allowing application installs from untrusted sources may be prompted to install apps they download through a link.

Since QR codes visually obfuscate links, an attacker may be able to trick an unsuspecting user into installing an Android application in this way. Thus, a natural question to ask is whether any of the apk links in our dataset point to malicious Android applications. To investigate this question, we first downloaded each of these apk files that was still accessible on the web, and then submitted it to the VirusTotal API for scanning. VirusTotal[5] is a subsidiary of Google that scans files and URLs using multiple antivirus scanners and website engines.

[4] https://developers.google.com/safe-browsing/
[5] https://www.virustotal.com/

Indeed, we do find instances of known Android malware among the apk links in our dataset. We attempted to download the apk files from a random sample of 4000 scans containing apk links. Of these, 2591 downloads were successful, with the rest failing due to 404 (not found) errors or connection timeouts. We submitted each of these applications to VirusTotal, receiving a result specifying the number of third-party virus scanners checked (generally 40-60) and the number of these that flagged the file. We investigate the VirusTotal report for each application that triggered more than 5 warnings. We find that 26 applications (1.0% of the 2591 apps we downloaded and scanned) are classified as explicit malware (e.g., Trojans). Another 45 applications (1.7%) are classified as Adware, and another 26 (1.0%) are classified as otherwise suspicious. None of these flagged applications appeared twice in the 4000 scans we considered.

We were surprised at the relatively high percentage of apk files flagged as malicious or suspicious. Possibly, some users scan QR codes with the intent of downloading applications to help them root their phones (such apps are considered malware by VirusTotal). For example, this video walks a user through the process of rooting their phone and shows a QR code that, when scanned, will download the rooting application: `https://www.youtube.com/watch?v=np18BC6B0OY` (at 1:35). This use case may account for the relatively high number of malicious apk files in our dataset.

We spot-checked some of the URLs pointing to malicious apk files against the Google Safe Browsing API. The API does not necessarily flag these URLs as malicious. This suggests that it may not be sufficient for a scanning app to check URLs against the Safe Browsing API or similar.

6. DISCUSSION AND FUTURE WORK

We conducted a systematic, in-depth analysis of barcode and QR code usage in the wild. Our results show that barcodes and QR codes, which can enable a physical object to communicate information to a mobile device via a visual channel, are widely used. We analyze that usage in depth. Overall, we believe that our results should be encouraging the researchers and industry practitioners developing new ways of leveraging barcodes and QR codes with mobile devices. We see strong evidence of emerging use cases in our dataset. Our results do, however, provide a cautionary note: while some users seem to frequently scan codes, other users seem to use their code scanner more as a novelty (with some users only scanning a single code).

We now turn to several lessons from our study, as well as recommendations to the designers of QR and barcode scanning applications.

Key Lessons. Stepping back, we summarize the key lessons our analyses reveal about the use cases for QR codes and the frequency with which real users encounter these use cases in practice. These lessons include:

- *QR codes are an effective way to reach some users, but many users are infrequent.* QR codes are still an active technology. However, their use is not universal or uniform: the top users in our dataset performed a stunning number of scans. Half of scans were performed by only 10% of devices, suggesting that this set of users is easily reached by QR codes (e.g., in marketing campaigns). However, many users scan only infrequently: almost one third of devices in our dataset scanned only once. Indeed, we find that despite a steady adoption rate by new devices, the rate of total scans levels out.
- *Web use cases dominate QR codes.* The majority (75%) of scans are of QR codes containing web URLs, suggesting that QR code use is dominated by the use case of quickly connecting users to websites. The popular domains appearing in QR codes do not correspond with domains that are popular on the web in general, with higher popularity of domains specific to QR codes (e.g., qrs.ly) and certain use cases (e.g., lottery).
- *Nevertheless, non-web uses are prevalent and varied.* Though non-web codes accounts for only about 25% of all scans (14% of QR code scans), the raw numbers of such scans are still significant. Moreover, the non-web use cases are distributed across a variety of different uses, including Wifi and one-time password setup codes, Bitcoin transactions, and other more niche uses cases. Thus, QR codes appear to be commonly used for a variety of emerging uses.
- *Some codes are intended for broadcast, while others are more limited use.* In comparing frequently and infrequently scanned QR codes, we observe that unique instances of some types of codes (e.g., Android app store URLs) are scanned more frequently than unique instances of other types (e.g., device pairing codes).
- *Scans and codes show no predictable geographic trends.* We observed no consistent geographic distributions of top codes: some are scanned only in one country, others are widely distributed. The cities where we observe the most scans are also geographically diverse. Our caveat about perspective of a single app still applies.
- *QR codes are not commonly used for malicious purposes, but users do encounter some malicious codes in practice.* Though we find that users in our dataset rarely encounter malicious QR codes, our dataset does several examples of malicious QR codes appearing in the wild, including URLs flagged as malicious by Google, links to Android applications containing malware, and a (possibly malicious) factory reset telephone code. Though these cases are rare, users may still encounter them, which informs our recommendations to creators of QR codes and scanning applications below.
- *Some code creators struggle to create correct QR codes.* We observed evidence of malformed QR codes of various types, including QR codes that containing JavaScript intended for a browser bookmarklet. In fact, we hypothesize that one of the top 100 most frequently scanned QR codes was created in error. System designers should not trust code creators to always create correct codes, and researchers should consider code creation as a possible point of failure.

Recommendations. Based on our findings, we make the following recommendations to the designers of QR and barcode scanning applications. Scan plans to take these recommendations into account in future version of the application.

Check for malicious or questionable URLs before automatically opening the link found in a QR code. Although we find that malicious URLs are uncommon, we nevertheless find some examples of malicious links in our dataset, including links to malicious Android applications, URLs marked as malware by the Google Safe Browsing API, and telephone codes that can factory reset some devices. We also find a significant number of links to adult sites, which, while not malicious, may not be appropriate for all users and in all contexts. Since users are unable to evaluate the content of a QR code visually, QR code scanning applications should be careful not to automatically load content, including a website's title, for potentially questionable links without first obtaining explicit user consent. We note that checking whether a link is malicious may not always be straightforward — for example, we observed that not all URLs leading to malicious `apk` files are flagged by the Safe Browsing API.

As wearable devices with the capability to quickly and automatically read QR codes, such as Google Glass, become more common, this recommendation will become even more critical. There have already been attacks against Google Glass that take advantage of automatic QR code reading, tricking Glass devices into joining unsafe Wifi networks [12].

Take steps to protect private information. Some codes contain private information, such as secrets used to authenticate to Wifi networks, initialize one-time passwords, or access Bitcoin private keys. These codes expose secrets to bystanders, who may be able to intentionally or unintentionally shoulder surf (especially if the bystander has a wearable devices such as Google Glass). Thus, designers of systems that use QR codes should consider the privacy needs of the data in the codes. They also need to consider the set of all potential scanning apps as part of the system's trusted computing base since, if the scanning applications are not trustworthy, the system's privacy properties may not be met. On the scanning side, perhaps the results of a scan of a QR code that contains certain classes of private information (such as private keys) should not be shown to users unless the user (rather than the device) has explicitly initiated the scan; this design would eliminate accidental shoulder surfing.

Users may face significant security risks if the ways in which scanning tools and system designers use QR codes don't match up well with users' expectations of the sensitivity of code data. While system designers should take steps behind the scenes, communicating a better model to users of which actions may be risky and which data may need to be protected could improve security significantly.

As a result of these findings, Scan plans for future releases to (1) add a user-friendly option to opt-out of data collection and to (2) emphasize their data collection policies directly in the user interface of the app.

Future work. We briefly outline directions for future work.

First, an additional investigation of QR code misuse that we may be able to conduct with our dataset is an exploration of QR code based attacks attempting to exploit QR and barcode scanning applications themselves (e.g., via input validation vulnerabilities and other attacks described in related work [7, 16]) or the website linked to in the QR code (e.g., via malicious query parameters). We did not study these exploit attempts in this paper because, for example, we do not know of a public repository of exploits against different scanning app + operating system configurations. A rigorous analysis might involve running each QR code through different combinations of scanning apps and operating systems, e.g., as was done for web browsers [23, 33].

We would also like to explore how much of the web linked from QR codes is not reachable by crawling the general web. The answer to this question has implications for any application that relies on the reachability of sites via a web crawl. For example, an application like Google Safe Browsing may find webpages to scan based on a crawl and might therefore miss websites linked only in the QR code based web. This gap would in turn pose challenges to QR code scanning applications attempting to test scanned URLs for safety.

This study focused on what we can learn about QR code use from our dataset of real-world scans. It would be valuable to extend our knowledge with a user study that more directly investigated both code creators' and scanners' motivations and experiences with creating and scanning QR codes — e.g., building on [32] to understand how frequently users scan codes in specific physical locations. A user study would also allow us to learn about non-users of QR and barcode scanning applications, a population that is inaccessible in our current dataset by definition.

Limitations. Finally, we mention several limitations of our dataset. First, while our dataset is large and and diverse enough that we believe it provides us with general information about the use of QR and barcodes, we are nevertheless limited to this single vantage point of one QR and barcode scanning application. Other applications may be popular among different user groups or in different regions with different behaviors. Similarly, because our dataset ends at a particular date, we expect that devices and codes appearing late in the dataset may be underrepresented; they may become more popular after the end of our dataset. As discussed above, while we use the term "user", we do not strictly have information about users but about devices, which may not map one-to-one onto the set of users.

7. CONCLUSION

One- and two-dimensional barcodes, including QR codes, present a convenient way to link physical objects to digital actions and have been widely adopted in both commercial and academic settings. In this paper, we have presented (to the best of our knowledge) the first in-depth study of QR and barcodes in the wild, leveraging a unique dataset of 87 million scans from users of Scan, a popular QR code scanning application.

In our analysis, we examined general use patterns of QR and barcodes in the wild, finding that QR codes dominate barcodes and that some users interact frequently with QR codes in the wild whereas other users scan only a single code. We find that a majority of scans contain web URLs, but we also identify a wide range of varied and emerging uses of QR codes, including device pairing and crypto currencies (e.g., Bitcoin). We also identify misuses of QR codes, both by code creators who create malformed codes, as well as potentially intentional malicious behaviors, including links to malware.

Our findings allow us to develop an informed understanding about the types of QR codes being created and how users interact with them in the wild. From these findings, we distill concrete recommendations for QR and barcode scanning applications (e.g., to protect private data and check for malicious URLs). The sheer number of users, scans, and use cases represented in our dataset should be encouraging to researchers and industry practitioners developing new ways of leveraging QR and barcodes with mobile devices.

Acknowledgements

We thank our shepherd, Mary Baker, and our anonymous reviewers for valuable feedback on an earlier version. We thank VirusTotal for providing access to premium services for the scanning of APK files. We also thank Adam Shostack for the idea of looking for rickrolls and Alexei Czeskis for earlier discussions. This work was supported in part by NSF Award CNS-0846065, by the Short-Dooley Endowed Career Development Professorship, and by the University of Washington Tech Policy Lab, which was founded with a gift from Microsoft in 2013.

References

[1] C. M. Baker, L. R. Milne, J. Scofield, C. L. Bennett, and R. E. Ladner. Tactile graphics with a voice: Using QR codes to access text in tactile graphics. In *Proceedings of the 16th International ACM SIGACCESS Conference on Computers & Accessibility*, 2014.

[2] A. J. Bernheim Brush, T. Combs Turner, M. A. Smith, and N. Gupta. Scanning objects in the wild: Assessing an object triggered information system. In *Proceedings of the 7th International Conference on Ubiquitous Computing (UbiComp)*, 2005.

[3] Bitcoin community. Wallet import format, 2014. `https://en.bitcoin.it/wiki/Wallet_import_format`.

[4] A. Cammozzo. TagMeNot, 2011. `http://tagmenot.info/`.

[5] D. Chaffey. The Who, Why and Where of using QR or action codes for marketing. Smart Insights, sep 2012. `http://www.smartinsights.com/mobile-marketing/qr-code-marketing/qr-codes-location-demographic-statistics/`.

[6] Comscore. QR Code Usage Among European Smartphone Owners Doubles Over Past Year, Sept. 2012. `http://www.comscore.com/Insights/Press-Releases/2012/9/QR-Code-Usage-Among-European-Smartphone-Owners-Doubles-Over-Past-Year`.

[7] A. Dabrowski, K. Krombholz, J. Ullrich, and E. R. Weippl. QR inception: Barcode-in-barcode attacks. In *Proceedings of the 4th ACM Workshop on Security and Privacy in Smartphones & Mobile Devices*, 2014.

[8] L. Deng and L. P. Cox. LiveCompare: Grocery bargain hunting through participatory sensing. In *Proceedings of the 10th Workshop on Mobile Computing Systems and Applications*, HotMobile '09, 2009.

[9] T. Denning, A. Borning, B. Friedman, B. T. Gill, T. Kohno, and W. H. Maisel. Patients, pacemakers, and implantable defibrillators: Human values and security for wireless implantable medical devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2010.

[10] Dogecoin Project. Dogecoin, 2014. `https://dogecoin.org/`.

[11] Google. Google Glass: Setting up Wi-Fi. `https://support.google.com/glass/answer/2725950?hl=en`.

[12] A. Greenberg. Google Glass Hacked With QR Code Photobombs. Forbes, July 2013. `http://www.forbes.com/sites/andygreenberg/2013/07/17/google-glass-hacked-with-qr-code-photobombs/`.

[13] E. Hayashi, B. Pendleton, F. Ozenc, and J. Hong. WebTicket: Account management using printable tokens. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2012.

[14] O. Israeli. 2013 QR Market Summary: Are QR Codes Dead or Alive? Visualead, dec 2013. `http://www.visualead.com/blog/2013-qr-market-summary-are-qr-codes-dead-or-alive/`.

[15] T.-W. Kan, C.-H. Teng, and W.-S. Chou. Applying qr code in augmented reality applications. In *Proceedings of the 8th International Conference on Virtual Reality Continuum and its Applications in Industry*, 2009.

[16] P. Kieseberg, M. Leithner, M. Mulazzani, L. Munroe, S. Schrittwieser, M. Sinha, and E. Weippl. QR code security. In *Proceedings of the International Conference on Advances in Mobile Computing and Multimedia*, 2010.

[17] Kirkmc. Create a QR code bookmarklet. Mac OS X Hints, Dec. 2012. `http://hints.macworld.com/article.php?story=2012112105493496&mode=print`.

[18] C. M. Li, P. Hu, and W. C. Lau. Demo: Authpaper - protecting paper-based documents/credentials using authenticated 2d barcodes. In *Proceedings of the 12th Annual International Conference on Mobile Systems, Applications, and Services*, 2014.

[19] F. Maggi, A. Frossi, S. Zanero, G. Stringhini, B. Stone-Gross, C. Kruegel, and G. Vigna. Two years of short urls internet measurement: Security threats and countermeasures. In *Proceedings of the 22nd International Conference on World Wide Web*, 2013.

[20] MarketingCharts. Data dive: QR codes, July 2013. `http://www.marketingcharts.com/online/data-dive-qr-codes-29525/`.

[21] J. McCune, A. Perrig, and M. Reiter. Seeing-is-believing: using camera phones for human-verifiable authentication. In *IEEE Symposium on Security and Privacy*, 2005.

[22] Misterjunky. Samsung S4 Dial pad Phone Codes Discussion. XDA Developers, Dec. 2013. `http://forum.xda-developers.com/showthread.php?t=2558791`.

[23] A. Moshchuk, T. Bragin, S. D. Gribble, and H. M. Levy. A crawler-based study of spyware on the web. In *Proceedings of the 13th Annual Network and Distributed Systems Security Symposium (NDSS 2006)*, 2006.

[24] S. Nakamoto. Bitcoin: A Peer-to-Peer Electronic Cash System, 2008. `https://bitcoin.org/bitcoin.pdf`.

[25] R. Naraine. Google testing login authentication via QR codes. ZDNet, 2012. `http://www.zdnet.com/blog/security/google-testing-login-authentication-via-qr-codes/10105`.

[26] L. Project. What is litecoin?, 2014. https://litecoin.org/.

[27] S. Robinson, J. S. Pearson, and M. Jones. A billion signposts: Repurposing barcodes for indoor navigation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2014.

[28] F. Roesner, D. Molnar, A. Moshchuk, T. Kohno, and H. J. Wang. World-Driven Access Control for Continuous Sensing Applications. In *ACM Conference on Computer and Communications Security*, 2014.

[29] J. Rouillard. Contextual qr codes. In *The Third International Multi-Conference on Computing in the Global Information Technology*, 2008.

[30] M. Simkin, A. Bulling, M. Fritz, and D. Schroeder. Ubic: Bridging the gap between digital cryptography and the physical world. In *ESORICS*, 2014.

[31] US Federal Communications Commission. What is a toll-free number and how does it work?, 2014. https://www.fcc.gov/guides/toll-free-numbers-and-how-they-work.

[32] T. Vidas, E. Owusu, S. Wang, C. Zeng, L. Cranor, and N. Christin. QRishing: The susceptibility of smartphone users to QR code phishing attacks. In *Proceedings of the 2013 Workshop on Usable Security (USEC)*, 2013.

[33] Y.-M. Wang, D. Beck, X. Jiang, R. Roussev, C. Verbowski, S. Chen, and S. King. Automated web patrol with Strider HoneyMonkeys: Finding web sites that exploit browser vulnerabilities. In *Proceedings of the 13th Annual Network and Distributed Systems Security Symposium (NDSS 2006)*, 2006.

[34] Wikipedia. Rickrolling. http://en.wikipedia.org/wiki/Rickrolling.

[35] H. Yao and D. Shin. Towards preventing QR code based attacks on android phone using security warnings. In *AsiaCCS*, 2013.