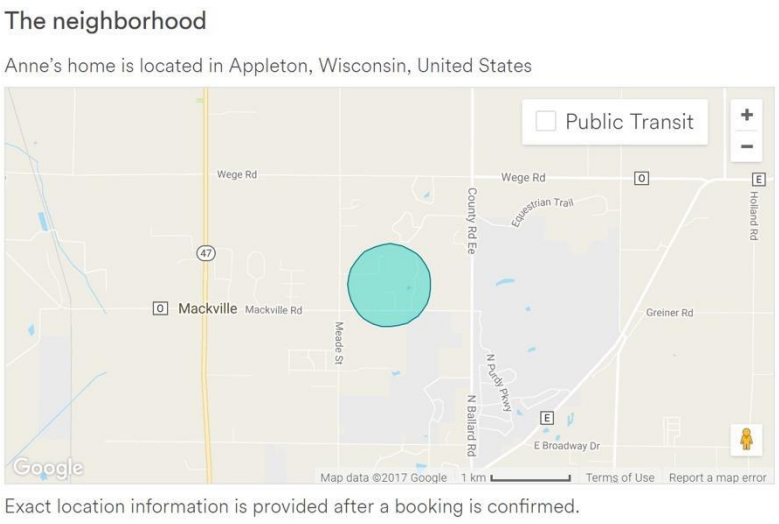
Szanto A, Mehta N. A Host of Troubles: Re-Identifying Airbnb Hosts Using Public Data. *Technology Science*. 2018100902. October 09, 2018. http://techscience.org/a/2018100902

A Host of Troubles: Re-Identifying Airbnb Hosts Using Public Data

Aron Szanto and Neel Mehta

Highlights

- We randomly sampled 693 Airbnb listings from Wisconsin from urban, suburban, and rural areas with varying population densities
- We introduce a method to probabilistically re-identify Airbnb hosts using public voter files from Wisconsin and the fuzzed location of Airbnb listings
- Despite Airbnb's efforts to protect personal data, we re-identified 94% of cases using the first name and town of resident living closest to the location provided by a listing



Example of an Airbnb listing that includes first name of the host, town of the listing, and approximate location

Abstract

Airbnb claims to make hosts anonymous by only providing their first name, town, and approximate location. In this paper, we introduce a method to probabilistically re-identify hosts by cross-referencing with public voter files, showing that Airbnb's anonymization method is vulnerable to attacks without access to specialized data. We survey Airbnb listings in nine Wisconsin localities of varying size and apply our algorithm to each, uncovering purportedly private host information.

Results summary: We find that, of the registered voters whose names match a host's first name and town, the voter who lives closest to the approximate location given by Airbnb is the true host in 94% of cases. This suggests that Airbnb's location fuzzing is insufficient to preserve the anonymity of their hosts. Further, we find that our method can uniquely re-identify the hosts of 40% of all listings in a random sample. Last, we detail how our methodology could provide regulators and law enforcement the tools to curb illicit renting activity but could also compromise the privacy of 1.2 million hosts.

Introduction

Increased access to the Internet and the rise of social networking have resulted in the emergence of the sharing economy. One of the largest players in this new space, Airbnb, brands itself as a "trusted community marketplace for people to list, discover, and book unique accommodations around the world"[3]. Hosts can list rooms, entire houses, or commercial properties on the site. It operates in nearly 200 cities, has over 3 million listings, and has served over 150 million guests. The company brands its product as an excellent method to monetize users' extra space. However, as opportunistic landlords use the platform as a way to circumvent hotel and rent control laws, Airbnb is under pressure by city and state governments to comply with local regulation regarding short-term rentals and to divulge information about hosts suspected to be in violation of the law.

The company fights such efforts by making it difficult to access data about the hosts on its platform, and through involvement in several lawsuits with major cities in the U.S. regarding data access and host behavior [18]. Private Airbnb listings hide identifying information about hosts, giving only their first name, town, approximate location, and occasionally pictures of the dwelling. Airbnb also censors any phone numbers or email addresses that users try to share [4]. As a result, regulatory officials do not have direct access to host information, curtailing their ability to enforce applicable housing law at scale. However, if such officials could re-identify these private listings on Airbnb without relying on data released by the company, they would not need to leverage legal channels to gain access to host data.

In this study, we considered a variety of listings from different parts of Wisconsin, randomly sampling from urban, suburban, and rural areas with varying population densities. For each, we attempted to re-identify the private host using publicly available data, including Wisconsin's public voter file, and the information in the Airbnb listing. We developed a proof of concept that could be used by city officials to uncover the identities of hosts suspected of

illegal activity, demonstrating also that our method leaves hosts open to the risk of theft and other harms if their listings can be re-identified.

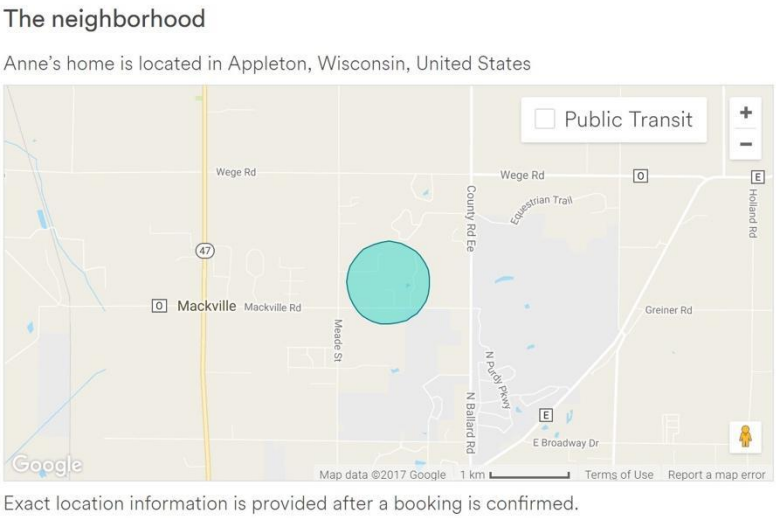


Figure 1. Example of an Airbnb listing's approximate location, as well as the first name of the host and the town the listing is in.

Background

An Airbnb listing, at a minimum, exposes an approximate latitude and longitude for a listing, the town the listing is in, the first name of the host, and an optional description section that may include a host's profession, hobbies, and educational background. We investigate whether this information is sufficient to reverse-engineer the host's identity and uniquely *re-identify* them.

Although to a first approximation the information exposed publicly on a listing may appear vague and anonymous, previous work shows that even simple demographic information like gender, date of birth, and zip code are sufficient to re-identify a large proportion of Americans [5]. Thus a first name and approximate address are good starting points for re-identification.

Re-identification of supposedly anonymized datasets has also been studied extensively. Researchers showed that cross-referencing parallel datasets, or multiple datasets with overlapping identifying data, can help re-identify people whose personal information was ostensibly private. For example, Narayanana and Shmatikov used an anonymized dataset released by Netflix regarding individuals' watching habits and cross-referenced it with an IMDB dataset to re-identify the users whose data was released by Netflix [6]. Sweeney used news reports to re-identify publicly released health records [7]. Our study used public voter data as a parallel dataset because each voter record contains the name, address, and town of each registered voter, allowing us to cross-reference with the information Airbnb provides.

Methods

Voter File

United States voter files include the name, address, and usually the demographic and contact information of every registered voter in a state, county, or other district. This includes *all* registered voters, including those who have not voted in a long time and those who have never voted. Each of the 50 states and Washington, D.C. maintain such a file, and these files are often available for free or for purchase from each state [24].

For our study, which emphasizes housing, geography, and population density, we use the state of Wisconsin as a case study because its population, population density, size, and home ownership rate are each close to the national average. Wisconsin has the 25th highest population density of the 50 states [9], is the 20th largest by population, and is the 23rd largest state by area [10]. Wisconsin also features a diverse mix of urban areas (such as Milwaukee, Madison, and Green Bay), suburbs (such as Sheboygan, La Crosse, and Wauwatosa), and rural areas (such as Portage, Reedsburg, and Elkhorn) [11]. Additionally, Wisconsin's home ownership rate is 67.8% [12], similar to the United States' overall rate of 63.5% [13].

Taken together, these facts suggest that Wisconsin is a reasonably representative sample of the United States for studies that focus on housing, geography, and population density.

Airbnb Listings

Airbnb's API is not publicly documented, so we used an open-source API wrapper to download metadata about Airbnb listings in Wisconsin [14]. Given a rough search area, e.g. Milwaukee, WI, and date range, e.g. January 1st to January 8th, the API returned a list of all Airbnb listings in that general area available within that date range.

The API revealed a variety of metadata about each listing, including:

- The listing's name (e.g. "Small Room in Milwaukee");
- The town the house is in;
- A fuzzed location of the house, given as a (latitude, longitude) pair;
- The host's first name;
- URLs of pictures of the house;
- A link to the listing on airbnb.com, which let us find the listing's description, reviews, and other details.

Airbnb shows users a shaded circle centered around the fuzzed location of the listing, as seen in Figure 1. After reviewing several listings from around the US, we found that the fuzzed location and circle are static, i.e., they are the same no matter when or where one views the listing. Moreover, the radius of the circle was very small, with an average of 0.284 miles and a maximum of 0.344 miles, and was uncorrelated with the population density of the locality of the listing. We determined this value by randomly sampling Airbnb listings in 10 US cities, finding the latitude and longitude of the center of each listing's circle. Then, using Google Maps, we found a landmark on the circumference of each circle, as seen in Figure 2. We then used Vincenty's algorithm [16] — discussed in detail in Appendix A — to find the distance between the circle's center and the landmark on the circumference, yielding the circle's radius.

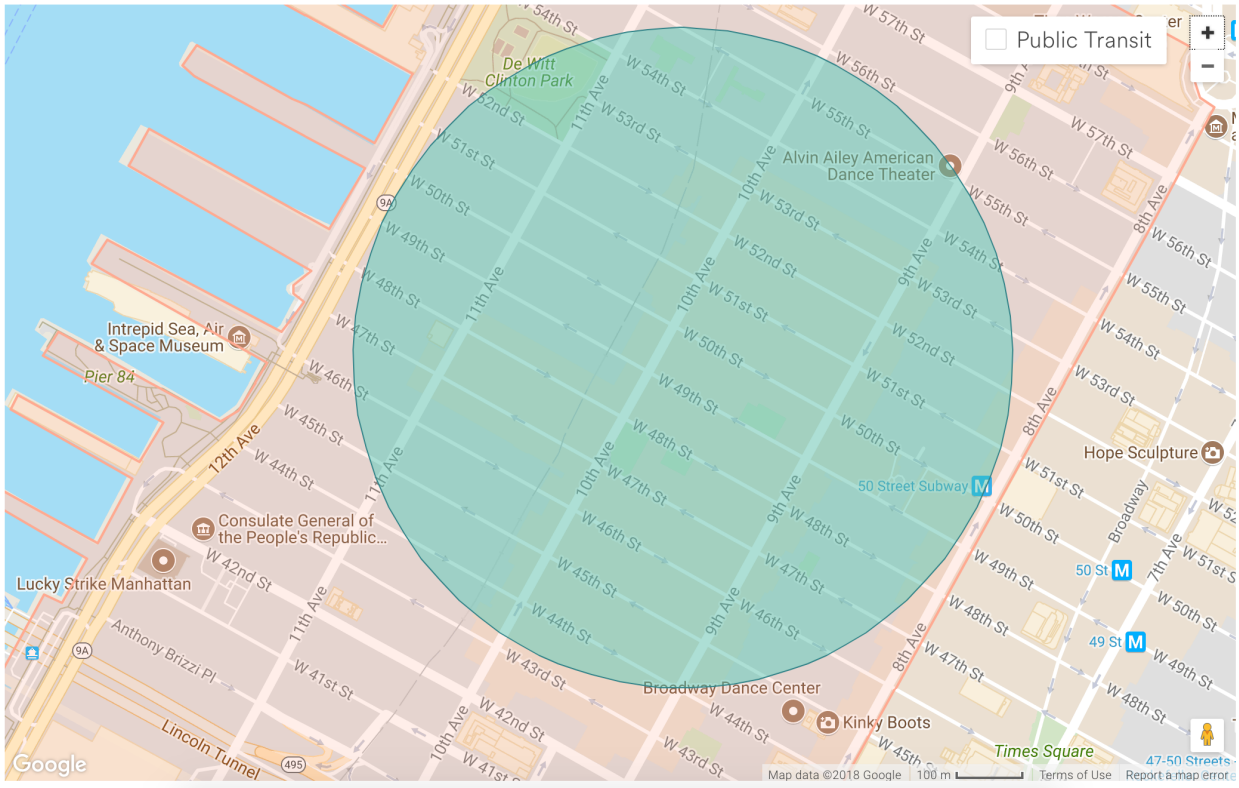


Figure 2. Example of calculating the radius of a circle drawn around an Airbnb listing in New York City. The Alvin Ailey American Dance Theater lies exactly on the circumference of the circle, so computing the distance from the theater to the circle's center (which can be found with an API) yields the circle's radius.

We gathered a full list of listings in nine selected Wisconsin cities, shown in

Table 1 and mapped in Figure 3 [15]. We chose:

- Three large cities (Milwaukee, Madison, and Green Bay);

- Three mid-sized cities with populations near 50,000 (La Crosse, Sheboygan, and Wauwatosa);
- Three small cities with populations near 10,000 (Portage, Elkhorn, and Reedsburg).

City	2010 Population	Wisconsin city pop. rank
Milwaukee	594,833	1
Madison	233,209	2
Green Bay	104,057	3
La Crosse	51,320	12
Sheboygan	49,288	13
Wauwatosa	46,396	14
Portage	10,324	84
Elkhorn	10,084	85
Reedsburg	10,014	86

Table 1. Selected Wisconsin cities and populations.

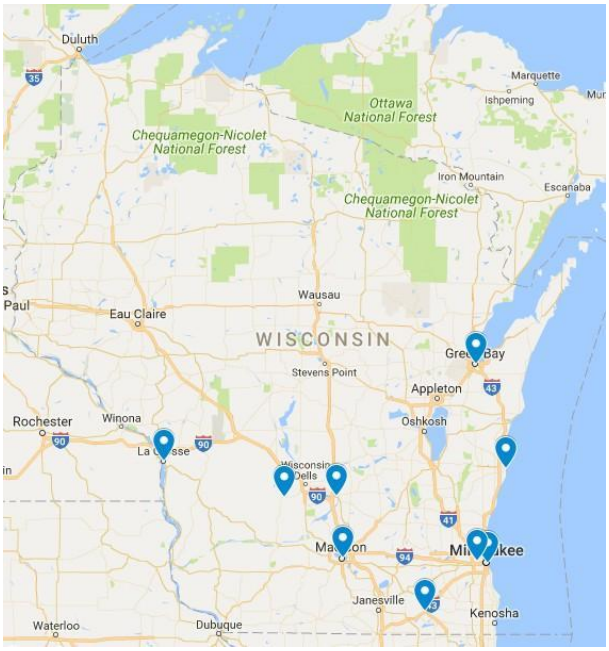


Figure 3. Map of selected Wisconsin cities. From west to east: La Crosse, Reedsburg, Portage, Madison, Elkhorn, Green Bay, Wauwatosa, Milwaukee, Sheboygan.

This allows us to compare the efficacy of our algorithm across different types of cities.

We randomly chose listings available between July 1st and July 2nd, 2017, which was roughly two months after the date of experimentation (late April 2017). We report 693 listings, as shown in Table 2.

City	# Listings
Milwaukee	117
Madison	127
Green Bay	87
La Crosse	25
Sheboygan	32
Wauwatosa	91
Portage	149
Elkhorn	48
Reedsburg	23
Total	**693**

Table 2. Number of Airbnb listings available from 7/1/17 to 7/2/17 for selected Wisconsin cities.

Algorithmic Re-identification

For each town, we sampled a random set of listings in the town, dropping expired listings and those removed by the host.

We then attempted to algorithmically re-identify the host and house of each sampled listing. Suppose we had an Airbnb listing A_i which provided the fuzzed location (Φ_i, L_i), where Φ_i is the latitude and L_i is the longitude. We filtered the Wisconsin voter records to only include voters with the same first name and hometown as the listing's host. We call these voters *candidates*.

For each candidate C_k, we ran their voter registration address through the Microsoft Bing reverse geocoder, which mapped the address to a latitude/longitude pair (Φ_k, L_k). We found the distance between the voter's house address and the Airbnb listing's fuzzed location. To do this, we used Vincenty's algorithm [16] to find the geodesic distance $s_{i,k}$ between (Φ_i, L_i) and (Φ_k, L_k). Our implementation of Vincenty's algorithm is discussed in more detail in Appendix A.

We then ranked all the candidates for the listing by the geodesic distance $s_{i,k}$ from their house to the fuzzed location. We hypothesized that the candidate with the lowest $s_{i,k}$ was the most likely to own the house.

Validation of Re-identifications

We proceeded to test the validity of our algorithm's predictions though three steps.

> **Step 1.** For each sampled listing, we found every registered voter whose first name and city matched the listing. These voters were called candidates.
>
> **Step 2.** For each sampled listing, we chose the 10 candidates closest to the fuzzed location; we call these finalists. (If there were fewer than 10 candidates, all became finalists.) We found that choosing the top 10 was sufficient to ensure correctness.
>
> **Step 3.** We then attempted to manually determine which of the finalists was most likely to be the real host of the listing. We split our methods into two groups: personal matching techniques and property matching techniques. The personal matching techniques included:

1. Using Google Reverse Image Search on the host's profile picture. Many times, hosts re-used that picture on other websites, such as on a LinkedIn profile, which revealed their last name and other demographics.

2. Searching LinkedIn or Google for the host's name, hometown, and this information often helped narrow down the list of potential matched voters, as some hosts' Airbnb profiles would provide identifying information such as their alma mater or job.

The property matching techniques included:

1. Using Google Maps Satellite View or Street View to visually compare a voter's house to pictures provided in the listing. Figures 5 and 6 provide an example of visual re-identification.

2. Using reviews and descriptions to infer the house's location. For instance, a review on one listing said that the house was "past three mailboxes, up the road, and the second house on the right". We looked up each of the matched houses on Google Maps and only found one that fit that criteria. In another case, reviews indicated that a listing was "blocks from the beach" and "next to a park and elementary school". Again, only one of the matched houses fit all of those criteria.

3. Leveraging the fact that many Airbnb listings were actual bed-and-breakfasts; hosts would just provide the names of the bed-and-breakfast companies (e.g. Sparrow Sun Wellness Bed and Breakfast). A simple Google search revealed the bed-and-breakfast's address and the owner's full name.

4. Using the situations in which listing owners included Google Maps screenshots of their properties, making them easy to re-identify.

After applying the personal and property identification techniques, we named a candidate a *unique re-identification* if one or both sets of identification techniques yielded a conclusive match. We note that there was never a case in which the two types of identification techniques pointed to two different candidates, i.e., there was never disagreement. Indeed, nearly all of the unique re-identifications were supported by both sets of techniques.

Results

We found unique re-identifications in 34 of the 84 Airbnb listings we tested, or 40.48%. That is, our algorithm was 40.48% successful in identifying a single individual who was very likely to own the house in a particular listing.

Most strikingly, of the 34 unique re-identifications, 32 of them represented the voter matching on name and town who lived closest to the fuzzed location. In other words, 94% of unique re-identifications could have been predicted just by finding the closest individual who matched the first name and town on the listing. We found that every individual uniquely re-identified was one of the 3 closest finalists, as seen in Figure .

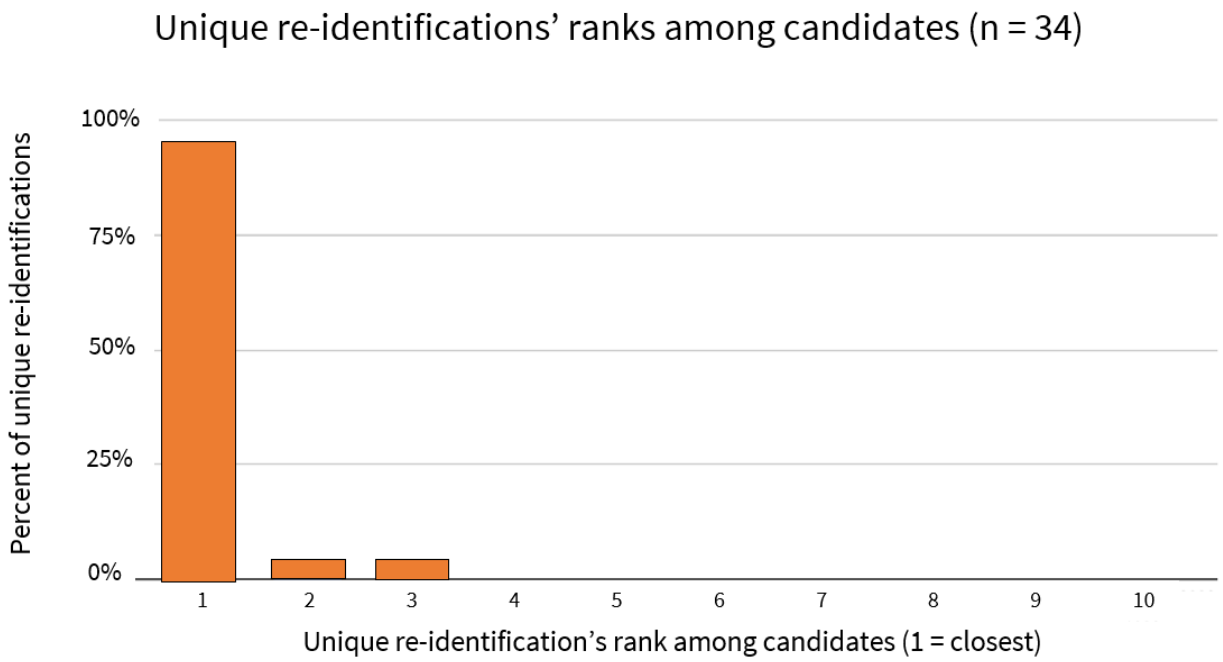


Figure 4. Assuming a match on first name and town, the voter who lives closest to the approximate location given by Airbnb is the true host 94% of the time.

This lends credence to our decision to name the top 10 closest candidates as finalists. Recall that we tested only the finalists to see if any of them were likely to be the owner of a particular listing. Thus, accidentally leaving the likely owner of a house out of the finalists pool could lead to false negatives. But since each unique re-identification was within the top 3 and almost all were the closest, choosing 10 candidates as finalists made it extremely unlikely that we would leave the likely owner out of the finalists pool. It gave significant buffer room, as no unique re-identification occupied a place between 4 and 10. As such, choosing 10 candidates gave us high confidence in our ability to correctly predict the true host.

Example Re-Identification

This example is based on a real listing that we uniquely re-identified, but we use pseudonyms and other fictitious details to preserve the privacy of the individual(s) involved. We denote details adapted to protect privacy with an asterisk (*).

In this example, consider a listing in Madison, Wisconsin owned by David*.



Figure 5. An image attached to one Airbnb listing.



Figure 6. We looked up one hypothesized address on Google Maps Street View. Comparing the images gives strong evidence that the listing is indeed at this address.

First, our algorithm found that there were 350* people named David in Madison, WI. That is, there were 350 candidates for this re-identification.

Our algorithm narrowed down these 350 candidates into 10 finalists, namely the 10 Davids* in Madison who lived closest to the fuzzed location of the Airbnb listing.

Those were David Adams, David Bell, David Cook, David Daniels, David Edwards, David Francis, David Gray, David Hodge, David Ian, and David Johnson*. The closest David, David

Gray, lived 0.09 miles from the fuzzed location (which was at the edge of one of the lakes in Madison), while the farthest, David Bell, was 0.50 miles away.

We hypothesized that the closest person was the most likely to be the host. We wanted to check if any of the algorithm's proposed finalists, and especially the finalist who lived closest to the fuzzed location, were the true owners of the house. Running our two sets of identification techniques, we found that for the personal case, Mr. Gray's Airbnb profile indicates that he attended Harvard University*. A Google search for "David Gray Harvard Madison"* brought up his LinkedIn page, which confirms that he attended Harvard. The LinkedIn profile picture also matches the Airbnb profile picture. For the property matching techniques, we searched Mr. Gray's address via Google Maps Street View and found that the house visually matched a picture provided in the listing. The listing also says that the house is on the water's edge, and Mr. Gray was the only David whose house was on the water. Last, an image of the property shows the view from the back porch, which includes a glimpse of the Wisconsin State Capitol building. Using Google Maps Satellite View, we found that the state capitol building was right across the water from this house and that the photographed view was plausibly taken from this house.

With these data points, we were able to name David Gray a unique re-identification. That is, we concluded that Mr. Gray almost certainly owns this property, validating the algorithm's prediction.

Cities

As Table 3 shows, our algorithm had varying success levels across cities. It was most successful in Sheboygan, with a 76% unique re-identification rate, but had limited success in La Crosse and Wauwatosa.

City	Listings tested	Unique re-identification found	% Unique re-identification
Milwaukee	8	5	63%
Madison	13	2	15%
Green Bay	9	2	22%
La Crosse	5	0	0%
Sheboygan	17	13	76%

Wauwatosa	2	0	0%
Portage	9	6	67%
Elkhorn	10	1	10%
Reedsburg	11	5	45%
Total	**84**	**34**	**40%**

Table 3. Unique re-identification success by city, sorted by city population (largest first)

We then grouped cities by their size, as seen in Table 4. The largest cities had the lowest success rate, while mid-sized cities saw the highest rate of likely matches.

City	Listings tested	Unique re-identifications	% Unique re-identifications
Small (≈ 10,000 people)	30	12	40%
Medium (≈ 50,000 people)	24	13	54%
Large (≥ 100,000 people)	30	9	30%

Table 4. Success of re-identification by city size

Distances

Of the 34 listings we identified as unique re-identifications, the associated properties were, on average, just 0.136 miles from the fuzzed locations provided by Airbnb. The distribution is shown in Figure 7, and full summary statistics are available in Table 5.

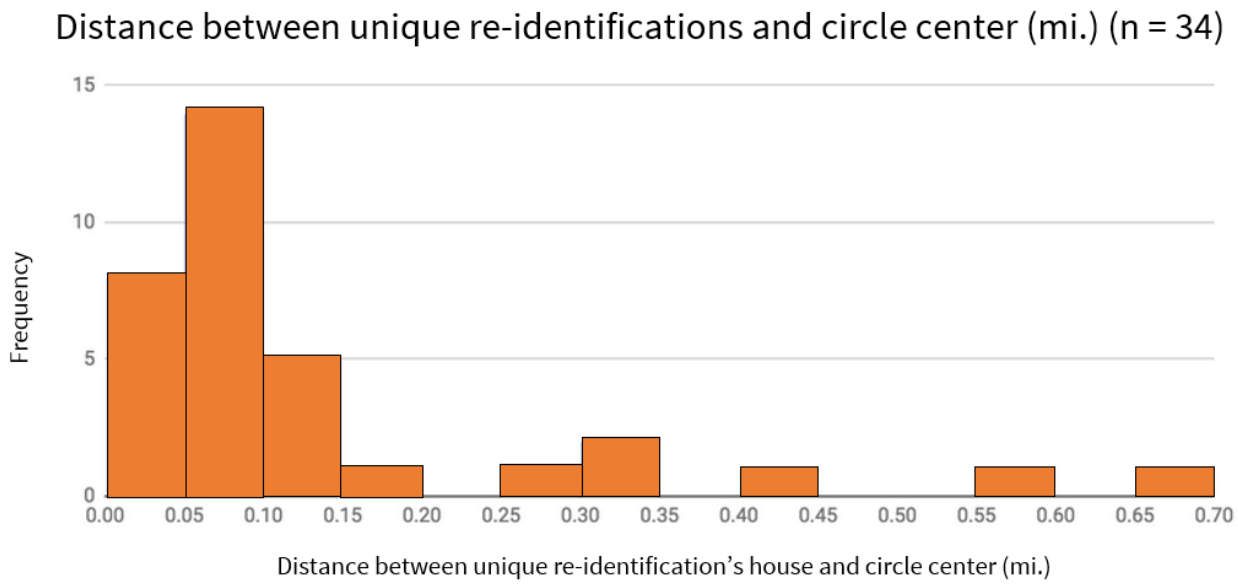


Figure 7. Distances between unique re-identifications and their Airbnb listings' fuzzed locations. On average, the likely houses are just 0.136 miles away from the fuzzed locations.

n	34
mean	0.136
std. dev.	0.157
min	0.018
median	0.076
max	0.684

Table 5. Summary of distances between uniquely re-identified properties and associated Airbnb fuzzed location.

Continuing Analysis

While our original study used listings collected in mid-2017, we repeated this study in April 2018 on four diverse, current Wisconsin listings. We present a case study on one such listing owned by a host named Sam* in Madison, including the other three in the appendix. We performed a unique re-identification for this listing as follows.

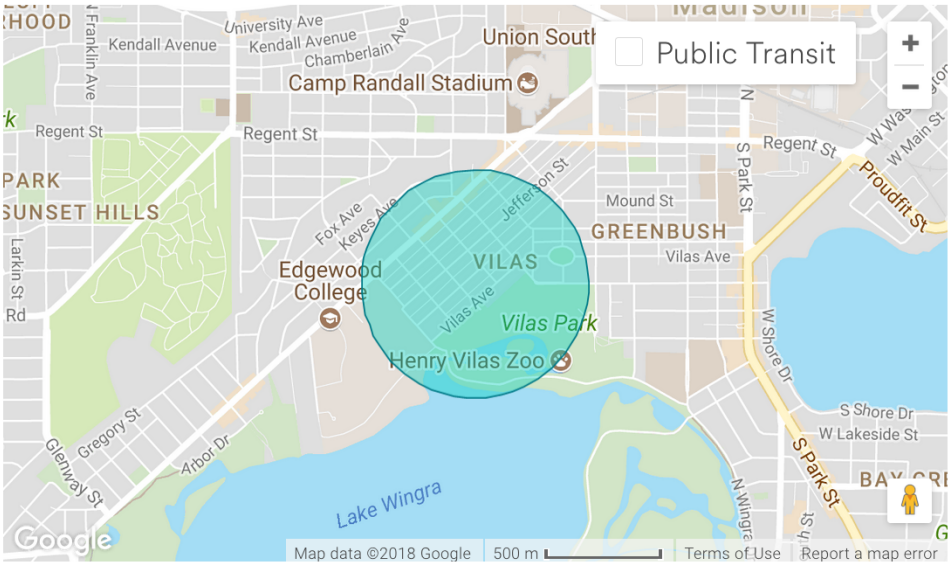


Figure 8. Fuzzed radius around the listing in question

Our algorithm found 667 Sams who live in Madison and are registered to vote in Wisconsin. Each Sam had an associated address from the Wisconsin voting records. We used reverse geolocation to find the latitude and longitude of each Sam's house, as well as the latitude and longitude of the Airbnb listing's radius. We found that Sam Reese* lived the closest to the center of the radius.

We confirmed that the listing belonged to Sam Reese using several methods:

- On her profile, Ms. Reese mentioned that she teaches geology at Quaker Ridge* College. A quick Google search for "Sam Reese Geology Madison" led to her faculty profile.
- The listing description includes, "It is very close to Block Park* (0.7 miles), Trader Joe's (4 blocks) and a variety of restaurants on Henry* Street. The zoo is nearby in Vilas Park." We used the distance of 4 blocks from Trader Joe's and proximity to Vilas Park to narrow down a small strip of area where the listing could be. Sam Reese was the only Sam whose house fell in this narrow strip.
- Google Maps Street View confirmed that this listing indeed matched Ms. Reese's house.



Figure 9. Listing image on Airbnb



Figure 10. Google Maps Street View of Sam Reese's house

Thus, we are confident in naming Sam Reese as a unique re-identification.

The center of the fuzzed listing's radius has latitude 43.062030906137714, latitude -89.41555717331067, per Airbnb. Ms. Reese's house is just 0.0617 miles (326 feet) away.

The combination of voting registration data and Airbnb approximate locations continues to be an effective tool for re-identification. Airbnb seems not to have made changes to its platform that would make this analysis materially different, with the location radius still under 0.3 miles for each listing in the new sample. We conclude that the results from the 2017 sample remain representative of the platform's privacy shortcomings in 2018.

Discussion

Our results suggest that despite Airbnb's efforts to keep personal data private, it is a relatively simple task to re-identify ostensibly anonymous users. Indeed, in 94% of cases, it suffices to find the voter matching on first name and town who lives closest to the location provided by a listing. These findings have important implications for regulators, users, and Airbnb itself.

Impact on Regulations

Our findings may be relevant to the ongoing regulatory battle between American cities and Airbnb. Cities object to Airbnb for facilitating renters' avoidance of hotel taxes and for supporting use of scarce housing supply for short-term rentals. Moreover, many of the properties are rent-controlled and are intended for low-income city residents, rather than for the operation of short-term housing rentals [18, 19]. Co-op boards and building management companies, meanwhile, have started to institute clauses in their lease agreements prohibiting short-term sublets to maintain the privacy and comfort of their tenants [23]. In New York and San Francisco, regulators and building management entities requested the names and addresses of Airbnb hosts, only to face difficulties in obtaining this data. Our results suggest that regulators and management alike could turn to an algorithm like ours to obtain this data themselves.

Below we explore two case studies for how our findings could help streamline the process of enforcing short-term rental regulation. In both cases, regulators require data regarding hosts whose listings are potentially illegal and have policies that they could enforce once they obtain the data. Our research could help these regulators with the missing middle step of acquiring the data without the cooperation of the company.

Hotel Registration in San Francisco

In 2017 the city of San Francisco won an injunction against Airbnb that requires all Airbnb hosts in the city to register with city regulators, providing information including name and address [18]. A victory for regulators, this ruling allowed them to put restrictions on hosts, such as limiting each host to one rental. These policies sought to prevent unscrupulous hosts from abusing rent-controlled housing units or from circumventing laws intended to protect hotel customers. However, this step forward came at the cost of an expensive and protracted legal battle, and it still allows for a "lag period during which illegal hosts can rent out homes before city officials identify them" [18].

Alternatively, city officials could use an algorithm like the one devised in this study to automatically re-identify hosts across the city. That is, this algorithm would help officials get much the same data that they are already looking for – namely, the names and addresses of Airbnb hosts – much faster than using legal or legislative means.

Illegal Landlords in New York

Cities could also use our algorithm to enforce regulation policies already on the books that they have difficulty enforcing. In New York, for example, some landlords use Airbnb to operate illegal hotels in rent-controlled housing units, to the detriment of the low-income residents for whom the housing is intended. Indeed, then-New York Attorney General Eric Schneiderman stated that two-thirds of the state's Airbnb listings are exploitative in this manner [20]. New York State passed a law in 2016 making it illegal to advertise a listing for housing for a term shorter than 30 days [21], a move widely seen as aimed at Airbnb.

Unlike San Francisco, New York found it fairly easy to pass short-term rental regulation, but despite the new rules, the city had difficulty enforcing the law due to the fact that neither Airbnb hosts nor their addresses were fully identifiable. New York regulators pressed Airbnb to expose the names and addresses of those individuals who were violating the law, but Airbnb refused, only allowing regulators to view anonymized, aggregated data by making an appointment at Airbnb offices in the city. Even then, city officials could only view and take notes but not leave with a copy of the data [19].

Given Airbnb's reluctance to divulge information about its hosts, our results suggest that cities such as New York might find it easier to crack down on hosts who attempt to sidestep the law by using an algorithm like ours to re-identify them.

Though a boon to regulators, it should be noted that our findings do not come without risk to individuals. Our methodology intersects the group of possible matches by location with the group of possible matches by name; this may return a group of people who match both criteria. Aggressive regulators might target an individual who meets both criteria but is not the true owner of the listing. Unless regulators ensure that they target only the correct individual, then "innocent bystanders" may be harmed due only to their circumstance of name and town of residence.

Last, we note that for ethical and institutional reasons, we did not contact any Airbnb hosts. Though this means that we do not claim positive proof of a correspondence, the evidence we find matching hosts to listings for our unique re-identifications is such that the probability of a false positive is vanishingly slim.

Other Outcomes of Re-Identification

Airbnb charges hosts a 3-5% commission on each transaction and guests a 6-12% fee for each rental [22]. As a result, Airbnb is incentivized to ensure that guests and hosts only interact

through the website. Indeed, the website implements stringent filters within its messaging service to prevent guests and hosts from exchanging contact information. However, our findings provide a natural way to circumvent these barriers by allowing would-be renters to find contact information for the listing elsewhere and to make an end-run around Airbnb's system. In addition to privacy risks, re-identification poses some more immediate risks for hosts. Instantly bookable homes on the platform, for instance, may be at higher risk for burglary. Clearly, the ability to re-identify the true owners and locations of Airbnb listings has wide-reaching and worrying effects.

Differences in Likelihood of Re-Identification

Our findings are differentially applicable to hosts who live in densely and sparsely populated areas. Because the probabilistic area that Airbnb provides has a fixed radius, hosts may find themselves either well-hidden or thoroughly exposed, depending on how many people live within that radius. For example, compare Figures 11 (showing a densely populated area) and 12 (showing a sparsely populated area). Assuming all other variables (e.g., population homogeneity) are equal between the two regions, it would be more difficult to re-identify the listing in Figure 11. Because the population density is so high in the area in Figure 11, there is a high probability of false positives, making it less likely — ceteris paribus — that the host could be re-identified. In contrast, the host in Figure 12 is more vulnerable — ceteris paribus — due to the relative sparsity of residential units in the location radius. Any positive match in the voter database for Figure 12 would likely be the correct match, in contrast with the many false positives that might be identified in the denser area of Figure 11.

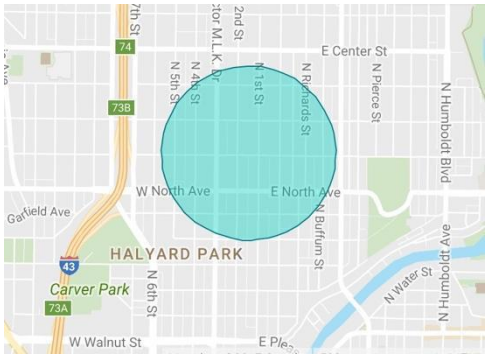


Figure 11. Re-identification would be more difficult here, ceteris paribus, due to the high population density within the radius.

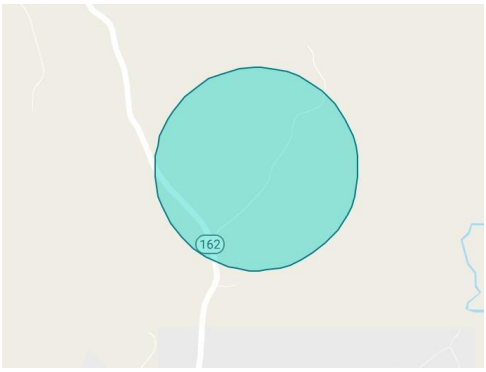


Figure 12. Re-identification would be fairly easy here, ceteris paribus, because there are few opportunities for false positives with such population sparsity.

To address the issue of population density, Airbnb might consider increasing the location radius to better protect hosts. However, while increasing the radius might decrease the risk of re-identification, the utility to prospective renters is significantly diminished. Many renters rely on the provided location to decide where they choose to stay, and greater uncertainty in location could make listings less appealing. While a renter in Coon Falls, Wisconsin may not notice a difference if the fuzzing radius was increased to one mile, a renter in New York City certainly would: a mile's radius could hide the difference between renting an apartment in downtown Manhattan and one in New Jersey!

Another potential approach could be to adjust the probability distribution of where the listing could fall within the location radius. As Figure 11 shows, the majority of listings were located close to the center of the fuzzed circle. Airbnb could make listings equally likely to appear anywhere in the circle, but again this would increase uncertainty of location and reduce the utility to potential renters.

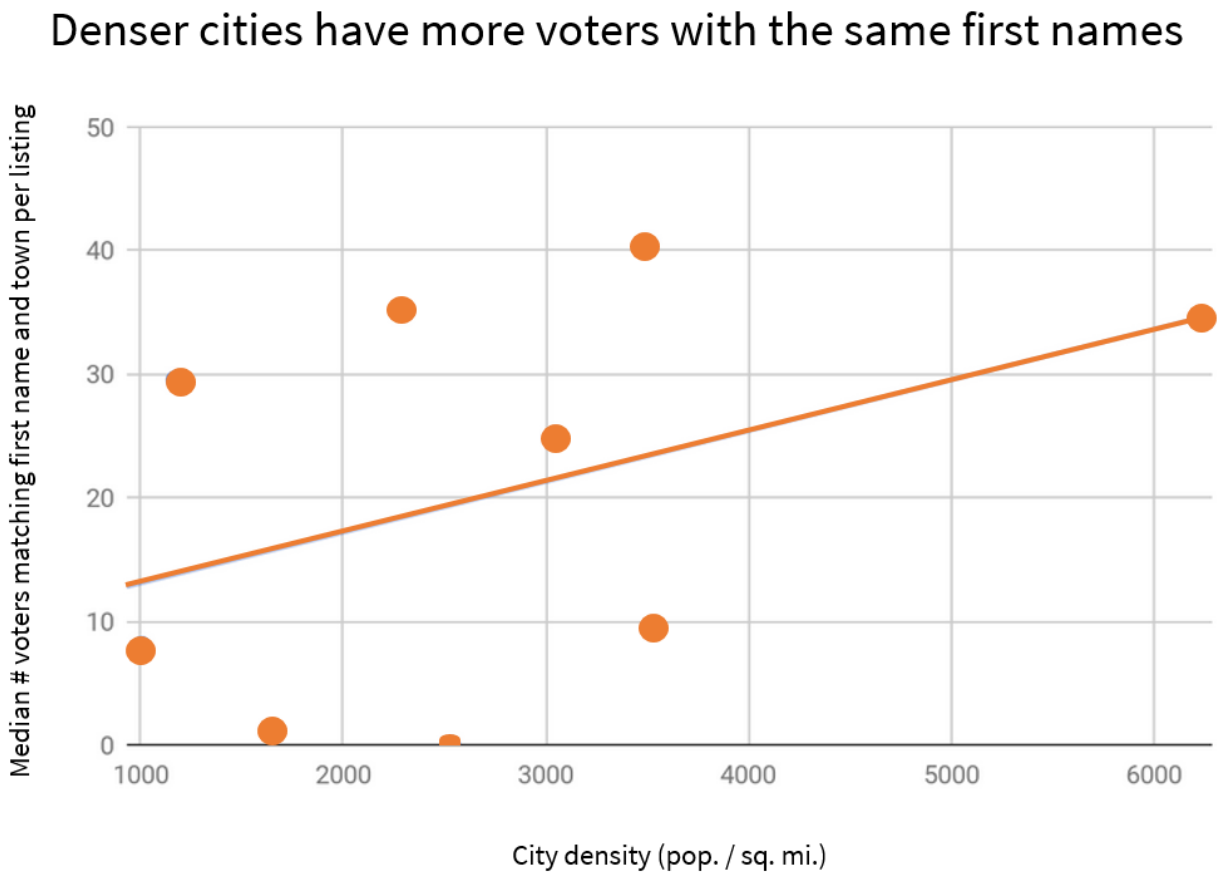


Figure 13. Denser cities had more voters with the same first names for the median listing, making re-identification harder.

This scenario highlights the tradeoff between privacy and utility: while increasing the security of users' anonymous information might decrease their risk of being identified, it makes Airbnb's service significantly less useful to users who have definite preferences as to their rental location. Regardless, if Airbnb were to adjust their service to address the vulnerabilities that we find, the first thing they might consider is to make the location radius dynamic, not static, and adjust the radius based on the population density in the area. Higher-density areas might require smaller radii, while lower-density areas could afford to have larger radii, which is especially important because fewer people will live in a circle of any given size in a lower-density area. Our analysis bore this trend out: Figure 13 demonstrates the significant effect of population density on the number of preliminary matches, i.e., the number of voters matching a listing's first name and town. In higher-density areas like Milwaukee, there could be as many as three dozen people, on average, who match these criteria in expectation.

Though differential population density might provide avenues for Airbnb to improve its service by adjusting location circle radii, it should be noted that other factors of population homogeneity may also affect the probability of successful re-identifications. For example,

high racial and age homogeneity in one locality may increase the number of people with the same first name, making re-identification harder still by allowing more people to pass through the name filter for a given town.

Airbnb might use other, more complex, methods to further reduce the risk of re-identification. The company could, for instance, use advanced techniques from computer vision and natural language processing, such as convolutional neural networks and entity recognition models, to prevent hosts from including photos of their house fronts or mentioning the names of their employers. While this may alleviate privacy concerns, reducing the amount of information available to guests about hosts may make it harder for trust to develop on the platform.

Shortcomings of Methods

Our re-identification algorithm was successful in re-identifying 40% of the sampled Airbnb listings, but this raises the question of why it was unsuccessful in the remaining 60%. We provide a few consistent patterns among the listings that our algorithm could not re-identify:

1. Our algorithm looked for all voters whose first names exactly matched the host's listed name. This worked fine for most names but failed in the case of nicknames. For instance, a man named Joseph might be entered in the voter records as Joseph but might go by Joe on Airbnb. After initially noticing the problem, we adjusted the algorithm to check if the host name was just contained within the voter name, instead of requiring an exact match. This would allow us to catch people named Samuel who go by Sam or people named Elizabeth who go by Beth. However, the case of Joe vs. Joseph would still cause our algorithm to fail. A simple extension to this project would use or build a database that maps nicknames to possible official names.

2. Furthermore, some couples listed two names on Airbnb, e.g. Jay and Ken. This would again break the name-matching algorithm. We attempted to rectify this by stripping any text that appears after "and", "&", or "+" (which would yield only the first of the couple's names), but this is fragile, and some couples might still slip by.

3. Some Airbnb hosts may not be registered to vote, meaning that our method would necessarily fail for them.

4. A few hosts had recently moved from other states or countries, so they were not included in the Wisconsin voter rolls. We imagine that this problem would be exacerbated in states with more in-migration, such as New York or California.

5. Several hosts did not live at the property they were listing on Airbnb; the properties could have been rental or investment properties instead of their primary residence. Extensions of this project could leverage public real estate records or commercial name/address databases to match properties to owners to solve this problem or the

previous one. For regulators in New York or San Francisco, this would be useful to crack down on hosts who list many properties on Airbnb.

6. In rural locations, Google Maps Street View was often unavailable, and Satellite View was not helpful because the houses were obscured by trees. In several such cases, we did not fully validate a re-identification, despite reasonably high confidence.

7. Hosts with common names were more difficult to re-identify because there are more matched voters in any given area. This made it more difficult to pinpoint which person the host was.

Future Studies

Our findings provide the basis for several areas of study. First, our re-identification algorithm relies on time-intensive human validation to confirm that the listing corresponds to the address listed on the voter record. We propose that our approach is conducive to execution via crowdsourcing platforms such as Amazon Mechanical Turk. This could render our approach highly scalable. Further, it may be possible to use advanced computer vision techniques such as convolutional neural networks to encode features of houses such that their publicly-listed images (on websites like Zillow and Google Street View) can be accurately matched against pictures provided in the Airbnb listing.

Since it's possible to choose an Airbnb listing and re-identify the host from voter registration data, it may be possible to do the converse. Given an individual's voter record, future work might show that it is feasible to perform a targeted attack to determine if that individual owns a listing on Airbnb. This may pose a considerable threat to targeted individuals' privacy and security.

Last, our re-identification algorithm might be improved by algorithmically cross-referencing with more data sources, which would provide stricter filtering and reduce false positives. Potential data sources include social media profiles, public property tax records, Zillow's database of homes, the Multiple Listing Service that allows realtors to find information about homes, and more.

References

1. Hamari, J., Sjoklint, M., & Ukkonen, A. (2015). The Sharing Economy: Why People Participate in Collaborative Consumption. *Journal of the Association for Information Science and Technology*, (August). http://doi.org/10.1002/asi.23552

2. Evans, J. (2016). Uber and Airbnb business models come under scrutiny. *Financial Times*, 40–42. Retrieved from https://www.ft.com/content/381e27ee-c685-11e6-8f29-9445cac8966f

3. Airbnb. (2017). About Us. Retrieved from https://www.airbnb.com/about/about-us

4. Tynan, D. (2011). The "new" Airbnb: Too little, too late? *ITworld*. Retrieved from https://www.itworld.com/article/2738422/it-management/the--new--airbnb--too-little--too-late-.html

5. Sweeney, L. (2000). Simple demographics often identify people uniquely. *Carnegie Mellon University, Data Privacy Working Paper 3. Pittsburgh 2000*, 1–34. Retrieved from http://dataprivacylab.org/projects/identifiability/paper1.pdf

6. Narayanan, A., & Shmatikov, V. (2008). Robust De-anonymization of Large Sparse Datasets. *Security and Privacy, 2008. SP 2008. IEEE Symposium on*, 111–125. Retrieved from https://www.cs.utexas.edu/~shmat/shmat_oak08netflix.pdf

7. Sweeney, L. (2013). Matching Known Patients to Health Records in Washington State Data. *SSRN Electronic Journal*. http://doi.org/10.2139/ssrn.2289850

8. Burris, A. L., & Fischer, E. A. (2016). The Help America Vote Act and Election Administration: Overview and Selected Issues for the 2016 Election. *Congressional Research Service*. Retrieved from https://fas.org/sgp/crs/misc/RS20898.pdf

9. United States Census Bureau. (2016). Resident Population Data. Retrieved from https://www.census.gov/2010census/data/apportionment-dens-text.php

10. United States Census Bureau. (2017). QuickFacts: Wisconsin. Retrieved from https://www.census.gov/2010census/data/apportionment-dens-text.php

11. Wisconsin Department of Tourism. (2017). Wisconsin Counties, Cities, and Regions. Retrieved from https://www.travelwisconsin.com/cities-regions

12. Trading Economics. (2018). Home Ownership Rate for Wisconsin. Retrieved from https://tradingeconomics.com/united-states/home-ownership-rate-for-wisconsin-percent-a-na-fed-data.html

13. United States Census Bureau. (2018). *Quarterly Residential Vacancies and Homeownership, Fourth Quarter 2017*. Retrieved from https://www.census.gov/housing/hvs/files/currenthvspress.pdf

14. Pham, S. (2017). airapi: Node.js client library for accessing Airbnb's API. Retrieved from https://www.npmjs.com/package/airapi

15. Wikipedia. (2017). List of municipalities in Wisconsin by population. Retrieved from https://en.wikipedia.org/wiki/List_of_municipalities_in_Wisconsin_by_population

16. Vincenty, T. (1975). Direct and Inverse Solutions of Geodesics on the Ellipsoid with Application of Nested Equations. *Survey Review*, *23*(176), 88–93. http://doi.org/10.1179/sre.1975.23.176.88

17. Veness, C. (2015). Vincenty solutions of geodesics on the ellipsoid. Retrieved from http://www.movable-type.co.uk/scripts/latlong-vincenty.html

18. Somerville, H., & Levine, D. (2017). Airbnb, San Francisco settle lawsuit over short-term rental law. Retrieved from https://www.reuters.com/article/us-airbnb-sanfrancisco-settlement/airbnb-san-francisco-settle-lawsuit-over-short-term-rental-law-idUSKBN17X254

19. Griswold, A. (2015). Airbnb has a really convoluted process for being "transparent" with New York City. *Quartz*. Retrieved from https://qz.com/571165/airbnb-is-sharing-more-data-but-in-a-really-convoluted-way/

20. Golding, B. (2014). Two-thirds of NYC's Airbnb rentals are illegal sublets. *New York Post*. Retrieved from http://nypost.com/2014/04/21/two-thirds-of-citys-airbnb-rentals-are-illegal-apartments-state/

21. Hinchliffe, E. (2017). No Vacancy: How Airbnb's New York City problem is just getting worse. *Mashable*. Retrieved from http://mashable.com/2017/04/03/airbnb-nyc-no-vacancy/9bd3Kpg7PqT

22. Airbnb. (2017). What are Airbnb Service Fees? Retrieved from https://www.airbnb.com/help/article/1857/what-are-airbnb-service-fees

23. Dakoff, H. (2016). For condo owners caught renting out units on Airbnb, fines can be steep. *Chicago Tribune*. Retrieved from http://www.chicagotribune.com/classified/realestate/ct-re-0522-condo-adviser-20160516-column.html

24. National Conference of State Legislatures (2016). It's a Presidential Election Year: Do You Know Where Your Voter Records Are? The Canvass, Issue 66. Retrieved from: http://www.ncsl.org/Documents/Elections/The_Canvass_February_2016_66.pdf

Authors

We thank Professor Latanya Sweeney for providing us with the Wisconsin statewide voter records, saving us considerable time and expense. We also thank Professor Sweeney, Jinyan Zang, Ji Su Yoo, the Data Privacy Lab, and the Technology Science Group for their guidance

and advice about our methodology and analysis. Emily Houlihan was vital in the creation of the proof of concept for this work as well as the first iteration of this report.

Aron Szanto is a technologist and researcher based in New York. Formerly Product Engineering Manager at MarketFactory, he now works in machine learning at Kensho. His prior work includes an artificial intelligence algorithm to detect fake news based on how an article travels through social media. He graduated in 2018 from Harvard College.

Neel Mehta earned his AB in Computer Science from Harvard in 2018. Since then, he has worked at Google as an Associate Product Manager.

Referring Editor: Latanya Sweeney

Citation

Szanto A, Mehta N. A Host of Troubles: Re-identifying Airbnb Hosts. *Technology Science*. 2018100902. October 09, 2018. http://techscience.org/a/2018100902

Data

Under review for data sharing classification.

Appendix

A. Vincenty's Algorithm

We use a slightly modified version of Vincenty's algorithm [16] as follow.

First let:

- a and b be the major and minor semi-axes of the Earth
- $f = (a - b)/a$ be the flattening
- ϕ_1, ϕ_2 be the geodetic latitudes
- L be the difference in longitude
- $\tan U_{1/2} = (1 - f) \cdot \tan \phi_{1/2}$
- $\cos U_{1/2} = 1/\sqrt{1 + \tan^2 U_{1/2}}$ and $\sin U_{1/2} = \tan U_{1/2} \cos U_{1/2}$
- $\lambda = L$, the first approximation

Then repeat the following steps until the change in λ is negligible:

$$\sin\sigma = \sqrt{(\cos U_2 \sin\lambda)^2 + (\cos U_1 \sin U_2 - \sin U_1 \cos U_2 \cos\lambda)^2}$$
$$\cos\sigma = \sin U_1 \sin U_2 + \cos U_1 \cos U_2 \cos\lambda$$
$$\sigma = \arctan(\sin\sigma / \cos\sigma)$$
$$\sin\alpha = \cos U_1 \cos U_2 \cdot \frac{\sin\lambda}{\sin\sigma}$$
$$\cos^2\alpha = 1 - \sin^2\alpha$$
$$\cos 2\sigma_m = \cos\sigma - \frac{2 \sin U_1 \sin U_2}{\cos^2\alpha}$$
$$C = (f/16)\cos^2\alpha\left(4 + f \cdot (4 - 3\cos^2\alpha)\right)$$
$$\lambda\prime = L + (1 - C) \cdot f \sin\alpha \cdot \left(\sigma + C \sin\sigma\left(\cos 2\sigma_m + C\cos\sigma(-1 + 2\cos^2 2\sigma_m)\right)\right)$$

Then:

$$u^2 = \cos^2\alpha \cdot \frac{a^2 - b^2}{b^2}$$
$$A = 1 + \frac{u^2}{16384}\left(4096 + u^2(-768 + u^2(320 - 175u^2)\right)$$
$$B = \frac{u^2}{1024}\left(256 + u^2\left(-128 + u^2(74 - 47u^2)\right)\right)$$
$$B\prime = B/4\left(\cos\sigma(-1 + 2\cos^2 2\sigma_m)\right) - B/6 \cdot \cos 2\sigma_m(-3 + 4\sin^2\sigma)(-3 + 4\cos^2 2\sigma_m)$$
$$\Delta\sigma = B \sin\sigma\left(\cos 2\sigma_m + B\prime)\right)\right)$$
$$s = bA(\sigma - \Delta\sigma)$$

Where, again, s is the distance between the two points in question.

B. 2018 Re-Identifications

Note: All names and personally identifying information have been modified. Modified personal information is marked with an asterisk (*).

A listing in Milwaukee was re-identified to Jane Smith*. The listing owner mentions that she runs Milwaukee's Theatre in the Park*, and her LinkedIn profile matches that description.



Photo of house from Airbnb listing



Google Maps Street View of Jane Smith's home

A listing by a Molly* in Green Bay was reidentified to Molly James* via Google Street View.

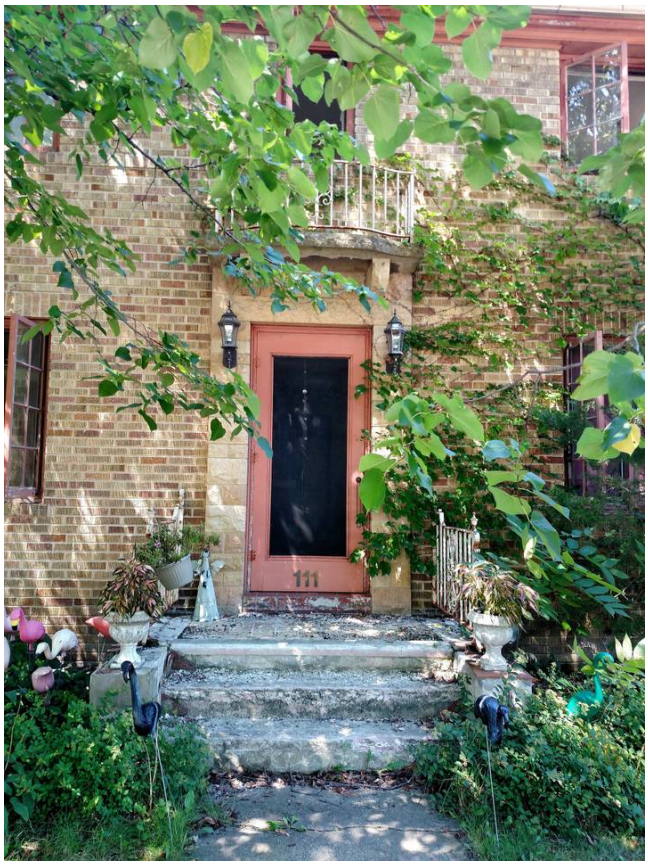


Photo of house from Airbnb listing



Google Maps Street View of Molly James's home

A listing by a Carmelita* in La Crosse was reidentified to Carmelita R. Burleson*, also via Google Street View.



Photo of house from Airbnb listing



Google Maps Street view of Camelita Burleson's home