An Adversarial Analysis of the Reidentifiability of the Heritage Health Prize Dataset

Arvind Narayanan*
Stanford University

May 27, 2011

Abstract

I analyze the reidentifiability of the Heritage Health Prize dataset taking into account the auxiliary information available online and offline to a present-day adversary. A key technique is identifying *providers*, which is useful both as an end in itself and as a stepping stone towards identifying members. My primary findings are: 1. Grouping providers based on shared members results in the formation of clusters which likely correspond to *hospitals*; 2. There is enough auxiliary information to identify most of these hospitals, and possibly also individual providers; 3. An adversary who has detailed information about a member's health conditions will be able to uniquely identify him or her; 4. While there are numerous websites where users can share reviews, health conditions, etc., their adoption is not currently high enough to serve as a source of auxiliary information for a large-scale member-reidentification attack.

I provide bounds on the efficacy of the methods I describe, but time constraints prevented me from attempting a more complete attack. To the best of my judgment, reidentification is within the realm of possibility; however, it is far from straightforward and will require algorithmic sophistication as well as sleuthing for auxiliary data. While identification of providers might be useful to contestants for improving predictive performance, large-scale reidentification of members—that has the potential to pose a threat to privacy and to the fidelity of the contest—appears unlikely to be feasible due to the paucity of auxiliary information.

*e-mail: arvindn@cs.utexas.edu; web: http://randomwalker.info/

1 The Heritage Health Prize

The Heritage Health Prize (HHP) is a machine-learning contest to "develop a predictive algorithm that can identify patients who will be admitted to the hospital within the next year, using historical claims data." The goal is to lower healthcare costs by predicting and preventing unnecessary hospitalizations. The two-year contest offers a USD 3 million Grand Prize as well as "Milestone Prizes".[1]

The contest is sponsored by the Heritage Provider Network (HPN), an umbrella organization of healthcare providers consisting of nine medical groups in Southern California.[2] The technical platform is provided by Kaggle Ltd., an Australia and U.S.-based startup that specializes in hosting such contests.[3] Kaggle has hosted 18 contests as of this writing, but the HHP has by far the largest purse and prestige.

As is the norm in machine-learning contests, "anonymized" or "deidentified" data is available for download. The deidentification was carried out by a team led by Khaled El-Emam, and used techniques such as generalization and suppression to in order to satisfy a set of k-anonymity–like criteria [EKA+11].

My involvement stems from my interaction with Kaggle. A previous Kaggle contest, the IJCNN social network challenge, asked contestants to predict missing edges in a social network graph. My team won the contest by deanonymizing the graph (which was derived from Flickr); deanonymization allowed us to simply "look up" the edges on Flickr [NSR11]. Due to this work and my previous research on deanonymization and privacy, I was asked to be on the HHP advisory board, and to analyze the reidentifiability of the contest data. I was not provided the data early enough for my report to have any bearing on the data release.

The data consists primarily of claims. Members (patients) and providers (physicians, labs, etc.) are identified by pseudonyms. In addition to the member and provider, each claim lists the diagnosis and procedure codes and several other attributes.[4] There is limited demographic information associated with each member: generalized age (decade) and sex.

The rest of this document, together with the Abstract above, constitute the report I prepared for HPN and Kaggle. Sections 2–5 elaborate on points 1–4 in the Abstract, and Section 6 presents my concluding thoughts including suggestions for improving deidentification on the basis of my results.

[1] `http://www.heritagehealthprize.com/`

[2] `http://www.heritageprovidernetwork.com/`

[3] `http://www.kaggle.com/`

[4] For a full list see `http://www.heritagehealthprize.com/c/hhp/Data`.

2 Providers and hospitals

The natural way to view the Claims data is a graph—specifically, a bipartite multigraph of members and providers. An edge connects a member with a provider, and represents a *visit*. Typically a visit corresponds to a single row (claim) in the table, but occasionally more than one. The other attributes such as Length of Stay and Diagnosis Code may be thought of as being attached to these edges, i.e., visits.

A graph is a data structure that lends itself well to large-scale deanonymization [NS09], provided that suitable auxiliary information is available. There are about 145,000 members and about 17,500 providers listed; this suggests that there is far more information per provider than there is per member. Similarly, auxiliary information about providers is more easily available than about members, and is more complete, as we will see in Sections 3 and 5. For these reasons, in this section I will focus on identifying providers. To do this, it will help to convert the bipartite graph into a graph of relationships between providers alone.

2.1 Clustering providers

Providers vary greatly in the number of members they are connected to in the member–provider graph. About a third have only one member, whereas the maximum member count is around 36,000. Figures 1 and 2 are two ways to visualize the number of members per provider. Providers break down into various specialties, shown in Figure 3, whereas Figure 4 shows the breakdown of member count per provider by specialty. Unsurpisingly, labs handle far more patients than physicians.

Let's call two providers *related* if they share at least one member. A crude but reasonably effective way to measure the degree of relatedness is the cosine similarity between the two sets of members that each is connected to.[5] This gives us a *weighted graph* between providers that will be a fundamental construct in much of the following analysis.

Of the $\sim$17,500 providers, $\sim$17,400 are related to at least one other provider, and $\sim$17,300 are in the giant connected component of the provider–provider graph. Of the 4,650 providers that have at least 10 members, all are in the giant component. Providers with at least 10 members are numerically in the minority, but they are responsible for 99.3% of all visits.

[5]The cosine similarity between two sets X and Y is $\frac{|X \cap Y|}{\sqrt{|X| \cdot |Y|}}$

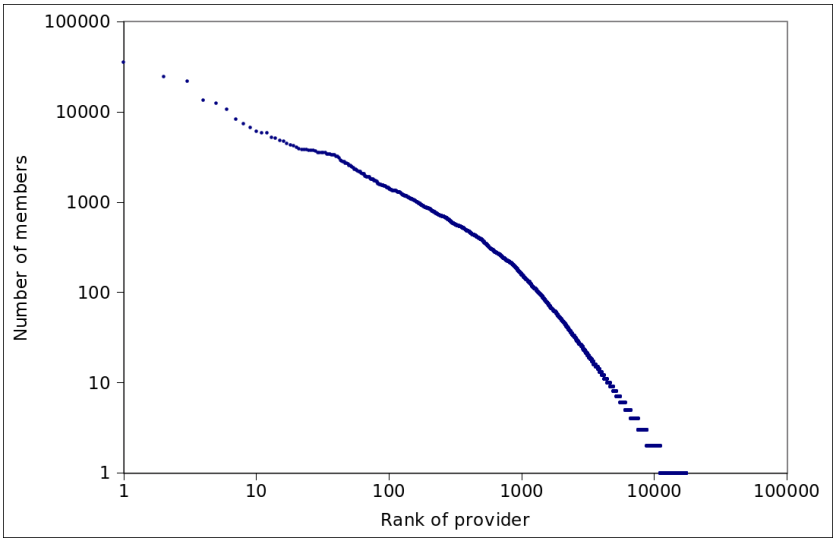

Figure 1: Rank of provider vs. number of members. Note the log-log scale.

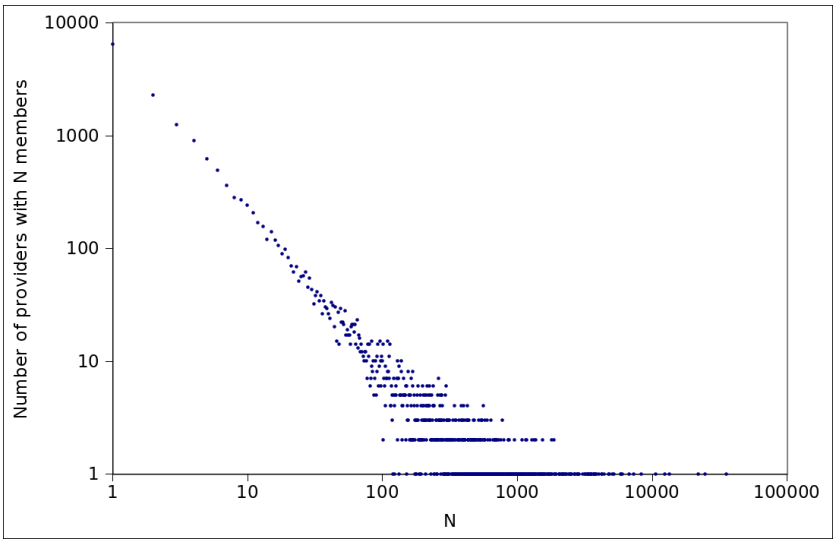

Figure 2: Number of providers with N members, for each N. Note the log-log scale.

Off-the-shelf clustering algorithms are rather dismal—they make rigid assumptions on the inputs and/or have a quadratic (or worse) running time. I therefore developed an algorithm tailored to the data. It works as follows: it starts

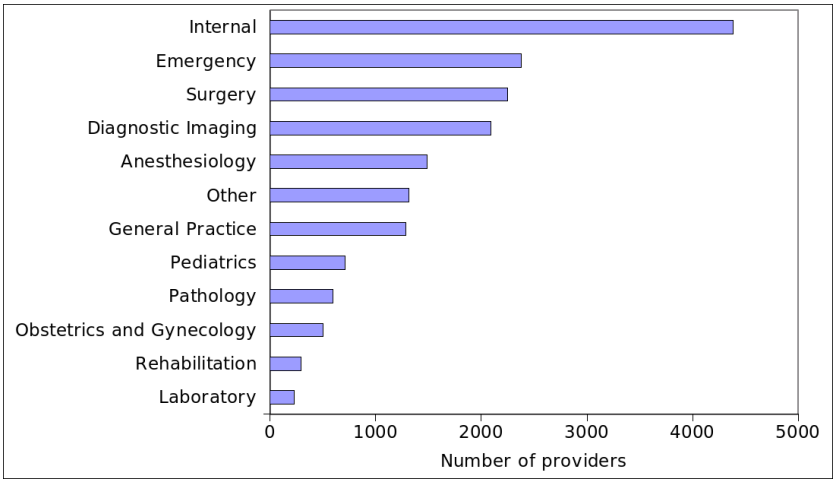

Figure 3: Number of providers, by specialty of provider

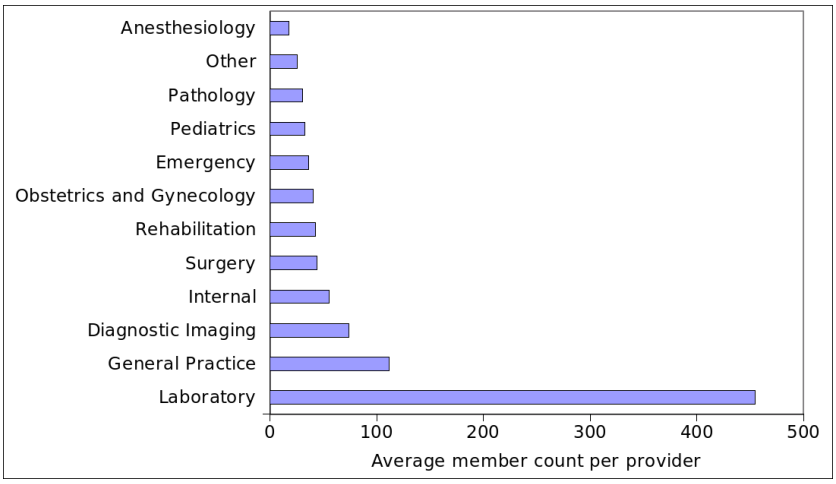

Figure 4: Members per provider, by specialty of provider

from a random point and accretively "grows" a cluster in a greedy manner. When the cohesion of the cluster (roughly, average edge weight) falls beneath a certain threshold, the cluster growth is stopped. Some overlap between clusters is permitted, but if the new cluster overlaps "too much" with an existing one, then it is discarded. Clusters of fewer than 10 providers are also discarded.

The algorithm found 180 clusters ranging in size from 10 to 190. Beyond this size the clusters are unable to maintain their cohesion. Together there are 4,400

providers that are part of a cluster, or about 25% of providers. This might seem like a small fraction but the reason will presently become clear.

The obvious question is what these clusters represent. Intuitively, there are two possible geographic scales at which clusters might form: individual hospitals and towns. The number and sizes of the clusters suggest that they are hospitals, but the clinching piece of evidence is the Place of Visit field. Only 12% of physicians who primarily see patients in the office were part of a cluster, whereas 42% of physicians who do so primarily in hospital settings were clustered.[6]

Figure 5 shows a more detailed view of this data, broken down by cluster. In the majority of clusters (70%), there are at least three times as many hospital visits as office visits; in the overall data, the ratio is reversed: there are over three times as many office visits as hospital visits. That said, there are a small number of large clusters (shaded area) with a hospital-to-office-visit ratio of under 1, and it is plausible that these represent small towns.

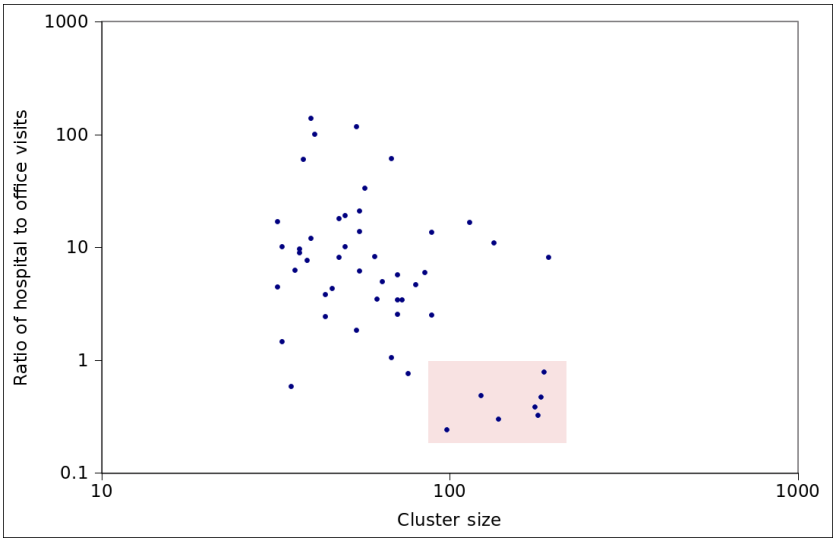

Figure 5: Clusters: size and composition. Note the log-log scale.

There are several ways in which the clustering algorithm can be improved. First, it has numerous parameters, and the size and nature of clusters seems to depend on the parameter choices. Second, one can incorporate medical knowledge/expertise about the nature of hospitals, for example, sharing of independent

[6]The 'Office' Place of Visit was treated as an office visit, the 'Inpatient Hospital', 'Outpatient Hospital' and 'Urgent Care' fields were treated as hospital visits, and the rest were ignored.

laboratories between hospitals. Third, currently the algorithm doesn't look at fields other than the member and provider, such as the primary care provider and vendor. I believe that a combination of these techniques can lead to identification of a higher fraction of hospitals, with lower false positive and false negative rates, as well as better identification of small towns.

2.2 Identifying hospitals

Due to the rich structure of the data, there are four different scales at which reidentification can happen: 1. cities/towns, which are large clusters or groups of clusters; 2. hospitals, which are individual clusters; 3. providers; and 4. members. These levels strongly interact with each other—for example, identifying a single provider who is known to practice in Bakersfield, CA will tell us which cluster of hospitals corresponds to Bakersfield. Now I will describe several potential techniques to identify hospitals, assuming the availability of suitable auxiliary information (which is described in Section 3).

- **Age.** If the average age of members that visit a particular hospital is unusually low, then maybe it is a pediatric hospital, or perhaps it is a hospital with a high proportion of pediatric doctors. If the average age is unusually high, then perhaps it is an assisted living facility.

- **Location.** As mentioned earlier cities/towns can potentially be isolated, but perhaps by observing relatedness on an even bigger scale we can estimate the geographic proximity between any two clusters. Armed with this information, identifying a small number of hospitals could lead to a cascading identification of all hospitals.

- **Hospital quality.** The connectedness between hospitals has a directionality: patients will get referred to the top hospitals from average hospitals if their condition doesn't improve, but the other way around happens much less frequently. A visit to two different hospitals for the same condition is indicative of a referral, and the Days Since First Claim field tells us which came first. By applying a pagerank-like algorithm, we can determine which clusters represent the top hospitals. Since the list of top hospitals is well-known, this gives us a relatively small set of possible matches.

- **Specialties.** The mix of specialties of doctors practicing in a given hospital might serve as a unique fingerprint. As a reminder, all of this depends on the fact that suitable auxiliary data exists, which is discussed in Section 3.

I was able to investigate the the first item in the above list, namely clusters whose members have unusually low ages on average. Figure 6 shows the mean member age for each of the top 50 clusters. There are a handful with mean age too low and a couple with mean age too high. I reran the clustering algorithm by restricting the original data to members who are in the 0–10 and 10–20 age groups. This resulted in 24 clusters of size at least 10. Some of these may be spurious, but I believe that at least the top 4—which have sizes 88, 62, 43 and 42 respectively—are truly clusters that represent pediatric hospitals or hospitals with a significant pediatric specialization.

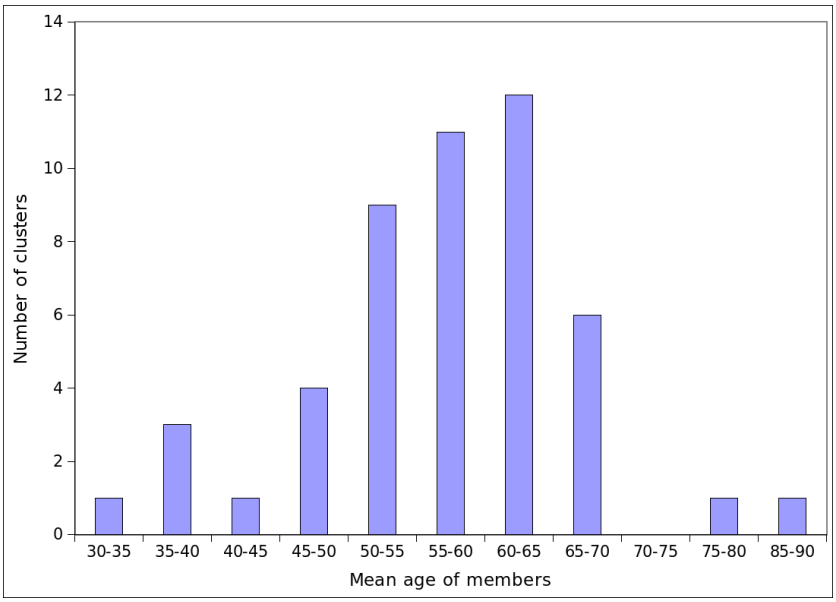

Figure 6: Age distribution of members associated with clusters

In the next Section I discuss how to determine which hospitals some of these clusters might be. But another question worth considering is how providers can be identified assuming that many or most hospitals have been identified. Physicians who practice at two are more locations, or who moved during the period represented in the dataset are prime candidates for reidentification. Given that specialties are also known, and considering that the number of physicians per hospital per specialty is fairly small, even a small amount of additional information such as average patient volume and referral patterns could be sufficient for identifying physicians. The State Inpatient Databases, discussed in Section 5.1 are an entirely different path to identifying providers.

3 Auxiliary information: providers

Now we are starting to reach the limit of what we can infer from the contest data alone, and we must go to external data sources to make further progress. In my investigations, I found three main types of useful external data on providers, although it is likely that there are others.

NPPES. The first is data published under the National Plan and Provider Enumeration System (NPPES).[7] It is a result of HIPAA which mandates the establishment of a unique identifiers for providers (the well-known National Provider Identifier, or NPI). In addition to identifiers, several pieces of information are collected from (self-reported by) each provider, and the aggregated dataset has been published regularly since 2007.

The current file is around 380MB compressed and contains 3.3 million providers. Of these, 380,000 list their Business Practice State as California. But at most 89,000 of these seem to be physicians (degree = M.D.); the others are dentists, psychologists, pharmacists, and various types of business entities. A typical record is shown in Appendix A.

There are many useful pieces of information here including business practice address, insurance plans accepted, and medical group. Given a hospital, it is easy to find the list of physicians who practice there. However, it must be noted that the majority of physicians do not report medical group information. It is likely that the quality of this dataset will improve with time (the NPPES website makes regular releases), but it is also possible that other data sources will help bridge the gap. Indeed, the remaining two that I will describe are useful mainly for that purpose.

Screen-scraping provider finder interfaces. Medical groups typically offer online services to the public including searching for a provider ("provider finder"). According to the HPN website, there are 9 affiliated medical groups.[8] I picked one of the larger ones, Regal Medical Group. As expected, RMG has a provider finder service.[9]

I was able to write a script to interact with the provider finder to extract all RMG-associated physicians in the following manner. Although a completely empty search doesn't work, it is possible to search for all providers in a city — the list of cities is available as a dropdown — and navigate all the pages of search results. This process yielded 1,500 providers in 170 cities.

[7] `https://nppes.cms.hhs.gov/`

[8] `http://www.heritageprovidernetwork.com/?p=medical-groups`

[9] `http://www.regalmed.com/members.cfm?m=phys`

The name and address are sufficient to link this dataset with the NPPES database. The provider finder has some useful additional fields like specialty and languages spoken. While I did not try other medical groups, I am confident that a similar process will work for all or most of them. After all, this is information that needs to be revealed to customers.

Hospital affiliation list. There is a list of hospitals utilized for HMO business by each California medical group.[10] HPN is affiliated with 54 hospitals in the list, but it is not clear what fraction of visits in the HHP dataset fall under HMO coverage. Interestingly, there is only one institution whose name suggests a pediatric focus: the Children's Hospital of Orange County (CHOC), which is apparently one of the busiest children's hospitals in the country.[11] Therefore I feel confident that CHOC corresponds to one of the 4 pediatric clusters mentioned in Section 2.1.

In conclusion, I have shown in this section that it is possible to obtain large-scale and accurate information about HPN-affiliated providers and hospitals. It is not quite complete, but in my judgement, an essentially complete list can be obtained with a little more effort and possibly some monetary expenditure. The adversary fundamentally has the advantage here because there isn't a strong norm or expectation of privacy for basic data about physicians and their affiliations. Finally, it is worth pointing out that other publicly available lists such as lists of top hospitals[12] and census data on population density and demographics[13] are also potentially useful for reidentification of providers and hospitals.

4 Member uniqueness

Threat model. There are 145,000 members in the dataset, and as pointed out by the deidentification team [EKA+11], the demographic information—age range and sex—is not nearly enough to identify members uniquely. But what if the adversary has some information about the member's medical history? That's what I will investigate in this section. Initially I will assume that nothing is known in terms of provider identities.

There are several people in our lives such as family, close friends and perhaps

[10] http://www.cattaneostroud.com/med_group_reports/21B-Web.pdf

[11] http://en.wikipedia.org/wiki/Children's_Hospital_of_Orange_County

[12] http://health.usnews.com/best-hospitals

[13] http://2010.census.gov/2010census/data/

neighbors who know something about our medical history. In the worst case, an individual who is potentially in the dataset might try to identify *their own data*, either out of curiosity or because an attacker bribed them to do so.

In addition to the demographic variables, the relevant attributes are Length of Stay, the Diagnosis Codes, the claim year and the Days Since First Claim field which roughly acts as proxy for date of visit in that the difference between two Days Since First Claim values equals the difference between the corresponding dates of visit.

There is no realistic generic model of auxiliary information in this situation—we can only have a meaningful model if we are considering a specific external dataset. Since we don't know what dataset the adversary might use, all we're left with is guesswork. In the deidentification paper, the authors make a specific, somewhat arbitrary set of assumptions. For example, they say "we assume that the adversary does not know the order of the quasi-identifier values. For example, the adversary would not know that the heart attack occurred before the broken arm...". I cannot fathom the reason for this. If the auxiliary information comes from online review sites, for example, detailed timeline information is very much available.

I believe that deriving reidentification probabilities based on such assumptions is a largely meaningless exercise. Nevertheless, I will perform such an analysis with a different set of assumptions, one that seems at least as realistic to me as the original. If nothing else, this should illustrate that mildly divergent assumptions can give very different results. Ultimately, the only relevant question is whether or not the auxiliary information is extensive enough; if it is, no deidentification procedure can possibly offer a meaningful level of protection.

I will assume that the adversary knows k of the diagnosis codes, for various values of k, and that he knows the approximate number of diagnosis codes (to within a factor of two). As for the timeline, I will assume a weak version is known to the adversary, namely the year but not the month of each visit.

There are numerous possibilities for what the adversary's auxiliary information could be. The timeline information could be more accurate or less accurate than in my model. The diagnosis codes known could be error-prone. On the other hand I have ignored the Length of Stay field, as well as lab results and procedures that I understand are planned for a future data release. I believe the choices above represent a reasonable middle ground. Also note that if the member's health conditions are known, there is no privacy risk to them from reidentification. Rather, the adversary's goal is to hack the contest or to find "seeds" for a large-scale attack.

An important issue is the fact that the released sample of members represents a fraction of HPN members, and a smaller fraction still of California residents (the adversary may not necessarily know which individuals are members of HPN). Any uniqueness metrics must account for this fact and estimate uniqueness *in the population* and not *in the sample*. For simplicity I will assume that the population size is 10 million, which is roughly equivalent to the adversary having some knowledge of the ZIP codes in which HPN-affiliated medical groups operate.

My methodology is as follows. For each k, I will calculate the n_k such that 50% of members are unique in a sample of size n_k if the adversary knows k of their diagnosis codes. This can be calculated as long as $n_k < \mu_k \cdot N_{sample}$ where μ_k is the fraction of members in the sample with k or more diagnosis codes. Making the assumption that the data is a random sample of the population, I will extrapolate this curve and find the k_{pop} such that at least 50% of members are unique in the entire population if k_{pop} diagnosis codes are known. Finally, I will calculate the fraction of members that have k_{pop} or more diagnosis codes.

In Figure 7, the blue line (bottom) represents the plot of k vs. n_k *without demographic information*. The curve is very close to a straight line, and I will generalize from this and assume that both of the lines in Figure 7 are straight line, and also that it remains a straight line if extrapolated to the right. One explanation for the straightness is a lack of correlation between different conditions. But that is not the only explanation; it is also possible that there is a small correlation but this is cancelled out because members with more conditions are also likely to have rarer conditions (which have more entropy).

The pink line (top) is the same plot but including demographic information. The pink line is the one that I'm actually after. Knowing the slope, I can extend the line and find k_{pop}, which is the k for which n_k exceeds $\mu_k \cdot N_{pop}$ (I'm assuming that μ_k for the sample is the same as for the population.) It turns out that $k_{pop} = 7$. In other words, roughly half of members with 7 or more diagnosis codes are unique if the adversary knows 7 of their diagnosis codes. This works out half of 25% or 12.5% of members.

In summary, I have shown that a significant fraction of members are vulnerable if the adversary knows demographic information and a sufficient number of diagnosis codes (and Year of Claim for each one). But if the adversary knows that a member is associated with a cluster that has been identified as a certain hospital, it changes the equation completely. Most clusters have only a few hundred members associated with them, which means that far less auxiliary information is required for reidentification. In particular, members who move to a different area during the data collection period are at high risk. The combination of hos-

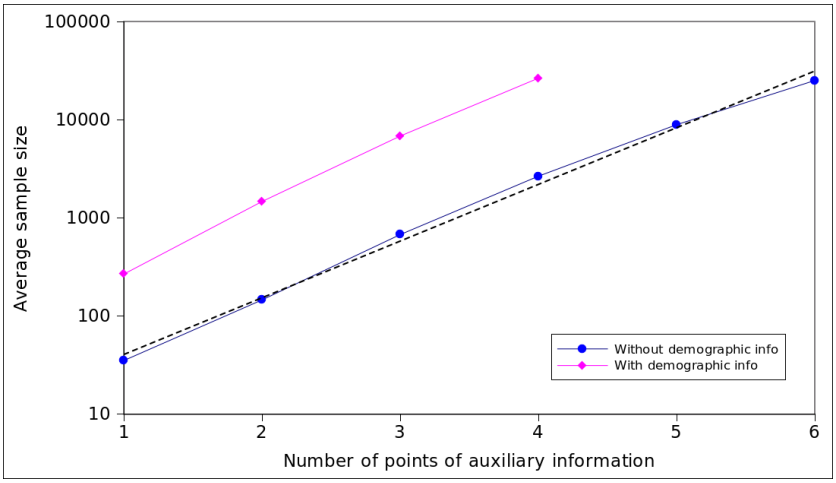

Figure 7: Average sample size for which 50% of members can be uniquely identified, as a function of amount of auxiliary information available to the adversary

pitals that they are associated with might be sufficient to identify them uniquely, provided that auxiliary information about their location (such as home address) is available.

5 Auxiliary information: members

5.1 State Inpatient Databases

By far the most interesting and important piece of auxiliary information on members seems to be data made available by The Agency for Healthcare Research and Quality (AHRQ), part of the U.S. Department of Health and Human Services. Under the Healthcare Cost and Utilization Project, the AHRQ makes available annual State Inpatient Databases (SIDs) from participating states.[14] Most states including California participate.

Although the SIDs constitute auxiliary information from the point of view of the HHP dataset, the SIDs themselves present a very serious reidentification risk. The AHRQ is cognizant of this risk, and have taken several non-technological measures to prevent it. First, obtaining the data requires (physically) signing a data-use agreement which prohibits reidentification. Second, completing an on-

[14] `http://www.hcup-us.ahrq.gov/sidoverview.jsp`

line Data Use Agreement Training Course is required. Third, there is a fee to obtain the data, although for some state databases including California, this is a token fee of $35. Finally, and the one that I find most interesting, is a requirement to describe the research project for which the data is going to be used.[15] If there is a human, especially one with domain knowledge, reviewing these applications, the last requirement can be very effective.

That said, if an adversary can manage to get his hands on the California SID, it can be a game-changer. In the deidentification document (Section 2.5.3—"Matching With Semi-Public Registry"), the authors say:

> if an individual can match the hospitalization records with the SID records, then the adversary can discover the exact month and year of birth of the member, their race, and their detailed diagnosis codes and procedures. This would more information than is disclosed and will therefore raise the re-identification risk.

Is this attack feasible? Unequivocally so. In the California SID databases, each patient is assigned a pseudonym with the explicit purpose of enabling tracking across hospitals and through time.[16] Further, the SID contains "relative dates" on a per-patient level, very similar to the Days Since First Claim field in the HPN dataset.[17] This means that for the purpose of matching these two databases, the adversary has far more auxiliary information than the analysis in the previous section; with accurate timeline information as well as diagnosis codes, I claim that most members can be matched in a straightforward manner.

To be clear, this does not mean that members can be identified, as the SID database contains only pseudonyms. But it does have two important consequences. The first is that when dealing with an adversary who has access to the SID databases, the HPN data can be assumed to contain year and month of birth and all the other attributes listed earlier (a full list of attributes is available on the AHRQ website.[18]) This must be borne in mind when reading the following subsection.

The second consequence is that once members are matched across the two databases, identifying providers becomes a lot easier. This is because there are

[15] The application form, which incorporates all these requirements, is available at `http://www.hcup-us.ahrq.gov/db/state/SIDSASDSEDD_Final.pdf`

[16] `http://www.hcup-us.ahrq.gov/db/vars/siddistnote.jsp?var=visitlink` and `http://www.hcup-us.ahrq.gov/db/vars/siddistnote.jsp?var=pnum_r`

[17] `http://www.hcup-us.ahrq.gov/db/vars/siddistnote.jsp?var=daystoevent`

[18] `http://www.hcup-us.ahrq.gov/db/state/siddist/sid_multivar.jsp`

pseudonyms for physicians in the SIDs, also with the explicit purpose of tracking physicians. Better, the hospital identifier—not a pseudonym, but an American Hospital Association identifier—is also provided with each record. This means that the adversary can bypass the whole process described in Section 2 of clustering physicians in the HHP dataset and matching them to hospitals.

5.2 Mining the Web

There are 3 main types of auxiliary information about individuals that an adversary might look for online: health conditions, hospitals visited, and physicians. Each is very useful for matching in a different stage of deanonymization, as we've seen in previous sections. Websites that make the identity of the user available, and not just a pseudonym, are particularly useful sources of auxiliary information.

Yelp appears to be the breakout winner in terms of the amount of publicly available auxiliary information about patient visits. This information comes from user reviews of both patients and hospitals. San Francisco General, a busy hospital, currently has 117 reviews while Alameda County Medical Center, a more obscure hospital has two reviews. Many hospitals from the affiliation list discussed earlier have no Yelp reviews. Active Yelp users typically make so much information public that they are identifiable for all practical purposes.

Relative to the number of hospital visits, however, the number of reviews is very low. The average internist sees about 1,000 patients per year[19] but the total number of reviews for a physician on Yelp rarely exceeds 10. At the hospital level, it is more informative to focus on trauma care centers since those are much more likely to elicit a review for the hospital itself rather than for an individual physician. A busy trauma care center such as SFGH experiences annual patient volumes of around 50,000, and as mentioned earlier, has 117 reviews. In both cases, only a fraction of a percent of visits seem to lead to reviews. Unsurprisingly, given the above numbers, the fraction of users who review *two* or more hospitals or providers seems vanishingly small.

Patientslikeme.com is an interesting case. There are 100,000 members and a good fraction have detailed medical information. The site collects and displays members' sex, age and city, but otherwise encourages anonymity. It is not clear if members have much anonymity against a determined adversary: in previous research I found that users often pick the same screen name across websites and this

[19] `http://gateway.nlm.nih.gov/MeetingAbstracts/ma?f=102275563.html`

can be used to match profiles [Nar08]. More anecdotal evidence about Patientslikeme members losing their anonymity can be found in a Wall Street Journal Article [AS10]. Another interesting fact about Patientslikeme is that the reported health conditions are available in machine-readable form (unlike all other websites surveyed here where such information is available only in the free-form text reviews that users write, if at all they choose to do so) although the mapping to the diagnosis codes in the HHP dataset may be far from straightforward.

Then there are numerous physician review sites: healthgrades.com, vitals.com, ratemds.com and UCompare to name a few. These sites are aimed at middle America rather than a young, urban, tech-forward population like Yelp is. The first three sites in fact seem much larger than Yelp in terms of number of medical reviews collected; however, they make little or no information available about users who provided the reviews, likely because the target demographic is not comfortable with this. Finally, sites such as Google Places and yellowpages.com contain a fair amount of hospital reviews but not reviews of individual providers. Users of Google Places are typically identifiable.

Auxiliary information about hospitals visited could come from an entirely different source: location-sharing websites. One benefit for the adversary of this type of data is that accurate timestamps are available, but an important caveat is that the user who checked in online might not be the patient.

Check-in data is easier to come by than reviews, as one might expect. The Cedars-Sinai Medical Center has been checked into 4000 times by 1200 users on Foursquare.[20] A list of "recent" check-ins for each location is available on the site but it doesn't appear that the full list can be crawled.

6 Concluding thoughts

In the preceding sections I have shown how hospitals and possibly providers can be identified by a determined adversary, and that while members who share information publicly about their hospital visits and health conditions are identifiable, not many currently do so.

Contest process. I find the AHRQ's data release procedure, described in the previous section, admirable and some of the steps potentially applicable to the HHP scenario. At the same time, I appreciate the importance of minimizing barriers to data download so that the participation rate does not suffer. But it might be

[20] `https://foursquare.com/venue/1959781`

possible to have the best of both worlds by dividing the process into two stages.

For the first stage, the current minimally intrusive process is retained, but the contestants don't get to download the full data. Instead, there are two possibilities. One is to release only a subset of the data. The other is to release a synthetic dataset based on the real data. Overstock.com recently announced this strategy for their contest.[21]

For the second stage, there are various possibilities, not mutually exclusive: require physically signing a data-use agreement and/or taking an online course, à la AHRQ; restrict the contest to the best performers from the first stage; and run the contest by having participants upload code to the server and obtain results, rather than download any data. The latter two strategies have again been announced by Overstock.

Auxiliary information and deidentification. Norms around public sharing of information are rapidly changing, and this is both a blessing and a curse from the standpoint of the HHP contest. It is a blessing because the contest data covers a period antecedent to the present time, and auxiliary data—especially location check-ins—gets progressively rarer as we go back in time. It is a curse because even user-generated data about *future* events might be useful auxiliary information. For example, some health conditions could be chronic. Another possibility is that within a few years, it might be the case that the majority of hospital visits are to be publicly disclosed via check-ins, at least in urban areas.

Perhaps the biggest saving grace, in my opinion, is that a random California resident has a probability of under half a percent of being in the HHP dataset. This means that the expected payoff for an individual from spending time trying to identify themself or a friend out of mere curiosity is very small. Trying to look for celebrities might be a realistic threat, and *large-scale* attacks are certainly well-motivated.

As for the deidentification process, what could be done differently? I gave some thought to methods for resisting reidentification of providers, such as suppressing some hospitals entirely so that the adversary's task of matching clusters to hospitals isn't 1-1. Ultimately, however, I believe these measures are futile, or at least, not worth the damage to the usefulness of the data.

It is much better to go after the adversary's weak point, which is the fact that even if the dataset can be augmented by matching it perfectly against the California SIDs, it doesn't seem possible to get to member identities based on public and

[21] `http://www.freedom-to-tinker.com/blog/aleecia/overstocks-1m-challenge`

quasi-public databases alone. Members who moved around might be vulnerable; for everyone else, some sort of self-reported online information about health or location appears necessary. A simple step could thus prove to be an effective way to forestall reidentification: write a script to scour the web for information about each member (using their real name, etc., *before* deidentification), and simply drop members from the dataset who have a habit of exposing personal information online. Based on current social trends, however, it may only be a matter of time before this strategy becomes untenable due to the ubiquity of self-reported auxiliary information online.

References

[AS10] Julia Angwin and Steve Stecklow. 'Scrapers' Dig Deep for Data on Web. http://online.wsj.com/article/SB10001424052748703358504575544381288117888.html, 2010.

[EKA+11] Khaled El Emam, Gunes Koru, Luk Arbuckle, Ben Eze, Lisa Gaudette, Emilio Neri, Sean Rose, Anthony Goldbloom, and Jonathan Gluck. The De-identification of the Heritage Health Prize Claims Data Set, 2011.

[Nar08] Arvind Narayanan. Lendingclub.com: A De-anonymization Walkthrough. http://33bits.org/2008/11/12/57/, 2008.

[NS09] Arvind Narayanan and Vitaly Shmatikov. De-anonymizing social networks. *IEEE Symp. Security and Privacy*, 0:173–187, 2009.

[NSR11] Arvind Narayanan, Elaine Shi, and Benjamin Rubinstein. Link prediction by deanonymization: How we won the kaggle social network challenge, 2011. http://arxiv.org/pdf/1102.4374v1; http://wp.me/pl3mS-bf.

Appendix A Example NPPES Record

NPI	1417950460
Provider Last Name (Legal Name)	GARCIA
Provider First Name	JOHN
Provider Middle Name	A
Provider Name Prefix Text	DR.
Provider Credential Text	M.D.
Provider First Line Business Mailing Address	PO BOX 129
Provider Business Mailing Address City Name	LOS ALAMOS
Provider Business Mailing Address State Name	NM
Provider Business Mailing Address Postal Code	875440129
Provider Business Mailing Address Telephone Number	5056619118
Provider Business Mailing Address Fax Number	5056619192
Provider First Line Business Practice Location Address	3917 WEST RD
Provider Second Line Business Practice Location Address	SUITE 139
Provider Business Practice Location Address City Name	LOS ALAMOS
Provider Business Practice Location Address State Name	NM
Provider Business Practice Location Address Postal Code	875442275
Provider Business Practice Location Address Telephone Number	5056619118
Provider Business Practice Location Address Fax Number	5056619192
Provider Enumeration Date	05/23/2005
Last Update Date	08/01/2008
Provider Gender Code	M
Healthcare Provider Taxonomy Code 1	207X00000X
Provider License Number 1	2003001211
Provider License Number State Code 1	MO
Healthcare Provider Primary Taxonomy Switch 1	Y
Other Provider Identifier 1	H81983
Other Provider Identifier 2	5729197
Other Provider Identifier Issuer 2	CIGNA
Other Provider Identifier 3	9274
Other Provider Identifier Issuer 3	EXCLUSIVE CHOICE
Other Provider Identifier 4	7135504
Other Provider Identifier Issuer 4	AETNA
Other Provider Identifier 5	179426
Other Provider Identifier Issuer 5	BLUE CROSS BLUE SHIELD
Other Provider Identifier 6	2305584
Other Provider Identifier Issuer 6	UHC
Other Provider Identifier 7	351240001
Other Provider Identifier Issuer 7	CIGNA DMERC
Other Provider Identifier 8	561510
Other Provider Identifier Issuer 8	HEALTHLINK
Other Provider Identifier 9	43511
Other Provider Identifier Issuer 9	HEALTHCARE USA
Is Sole Proprietor	N