ID:A Labs Finds 20 Million Improper Identity Manipulators

Dr. Stephen Coggeshall
Chief Technology Officer
ID Analytics

Abstract

In this paper, we will describe several insights derived from over 10 years of data-intensive analysis into the phenomenon of identity fraud. Specifically, this paper provides insight into the several categories of identity fraud, and describes tools and techniques used to detect and mitigate these behaviors. We also provide real-life examples and quantitative statistics for two specific use identity fraud modes: identity manipulation and the misuse of deceased identities

About ID Analytics

ID Analytics operates the ID Network, a cross-industry collaboration of data sharing for the purpose of identity fraud prevention. Large companies examine risk events, primarily new account openings or account changes, for the likelihood that it is an identity fraud. ID Analytics screens about a third of all U.S. credit card openings, new cell phone account applications, new check orders, and other applications for many other products and services such as auto loans, installment credit, mortgages. Additionally, we screen many account change events such as changes of address that can be instances of account takeover. All of these millions of events flow into and power the ID Network and give us unique real time visibility into the dynamics of identity fraud.

The ID Network contains more than 700 billion instances of personal identifying data, primarily names, addresses, SSNs, dates of birth, phone numbers, and emails, which allows visibility of more than 315 million unique people in the U.S. We have aggregated more than 1.7 billion consumer transactions that contain this personal identifying information (PII), including 2.9 million labeled fraud events, making the ID network the largest repository of labeled fraudulent account opening events in the world. The bulk of the 1.7 billion consumer events are applications for new account openings, which allows us to build tools for prevention of identity fraud around account origination.

We generally do not see the majority of card payment transactions as consumers purchase using credit cards, so we do not have substantial visibility into aberrant charging behavior that is indicative of a lost/stolen credit card account. Our visibility is at the account origination and during changes of PII, which are the places for classic identity fraud.

About ID:A Labs

ID:A Labs is the internal R&D group to ID Analytics, a company founded with the mission to prevent identity fraud. ID:A Labs reveals important trends in consumer behavior by examining identity use to help organizations better manage both the risk and opportunity of an individual consumer. ID:A Labs is a multidisciplinary group of mathematicians, computer scientists,

economists, financial experts, cognitive scientists and advisors from ID Analytics and other respected institutions.

ID:A Labs conducts research and analysis in the areas of identity fraud, credit risk, marketing and segmentation, authentication and identity proofing. The labs leverages the above-described ID Network™, the nation's first network of cross-industry consumer behavioral data. Backed by patented technology, world-class analytics, and the ID Network, ID:A Labs researches, analyzes and reports on developments in consumer behavior, identity- and credit-related issues, the regulatory landscape and innovations in analytics around modeling and machine learning.

What is Identity Fraud?

There are many different modes of fraud around identity and here we attempt to classify and describe the majority of the common types. First, we define identity fraud:

Identity fraud is the act of misrepresenting who (which person) you are in order to improperly obtain products or services.

Identity fraud means the misrepresentation of the particular person/individual you are—trying to represent yourself as a different individual. It **does not** include misrepresenting or hiding your behavioral characteristics such as past bad credit, but it **does** include pretending to be a different person to avoid past bad credit.

We make the distinction between what we call classic identity fraud and lost/stolen credit card fraud. While lost/stolen card fraud does include misrepresenting which person you are, this mode of identity fraud is limited to the use of that account only and is stopped with the closing of that account. Classic identity fraud allows the fraudster to cause more harm because with access to your personally identifiable information (PII) such as name, Social Security number (SSN), and date of birth, the fraudster has the ability to open many new accounts using the victim's persona. This paper deals primarily with this classic identity fraud.

Types of Identity Fraud

Lost/Stolen Account – A fraudster becomes aware of enough information around a specific account that allows him to impersonate you for account-level activity. This manifests itself as unauthorized transactions such as credit card charges or possibly unauthorized money transfers. This type of fraud is limited to the account level and as mentioned above, is not generally considered classic identity fraud.

Identity Theft – Also called true-name fraud, this is when a fraudster targets a specific real individual and assumes his persona. It typically manifests itself as unauthorized new account openings in the victim's name/identity. In this mode of fraud, the imposter is aware that he is improperly representing himself as a specific, real person.

Synthetic Identity Fraud – The fraudster fabricates a new and false identity that is not related to a real person and does not exist. The perpetrator will invent a set of PII and try to establish the existence of this invented identity. This may be accomplished through the repeated presentation of this collection of fictitious PII through various product applications and channels, with the intention of establishing this synthetic identity in the many existing credit-related

databases to be misused at a later time. There may be some unintended overlap to real PII but the core identity is artificial.

Identity Manipulation – In this mode, a fraudster will make what may be slight and/or subtle variations to his true PII in the hopes of confusing the system to avoid having the application be associated with his true identity. The fraudster may increment one digit of his SSN, or month, day or year of his date of birth, or interchange SSN digits or otherwise make manipulations of his real PII in the attempt to disallow the application process to link this presented application to the fraudster. Identity manipulators may apply for products using slight variations of their true identity to attempt to avoid past delinquent history. Sex offenders and illegal immigrants commit identity manipulation to live under aliases to avoid detection, while other identity manipulators seek to gain improper access to health care or government services and benefits.

The following table summarizes the main types of identity fraud, along with descriptions of the victims, the nature of the improper misrepresentation and some methods to catch the fraud.

Table 1. Summary of the Different Modes of Identity Fraud

	Identity Theft	Identity Manipulation	Synthetic Identity
Who is the victim	The person who's identity is being misused. The company providing the fraudulently obtained product or service is also a victim.	No consumer victim. The company providing the fraudulently obtained product or service is the victim.	No consumer victim. The company providing the fraudulently obtained product or service is the victim.
Nature of the misrepresentation	Typically SSN, name and date of birth belong to the victim, and the address and phone number are associated with the fraudster.	SSN, date of birth and/or name vary slightly from what's correct.	SSN, name and date of birth are fabricated or chosen randomly .
Signals to find/catch the fraud	Unusual activity around the SSN, name_dateofbirth combination (these belong to the victim), or address/phone (these are usually associated with the fraudster).	Look for systematic variations around the PII to differentiate from simple typos.	Closely examine the first instances of a PII assertion. Do they make sense? Are they related to other suspicious activity?
Summary	Victim's Core Identity	Fraudster's Core Identity	No Core Identity

ID Analytics Tools and Visibility into Identity Fraud

Above we described the data flow into the ID Network from which we can build a number of special tools, services, and perform specific analysis into the dynamics of identity fraud. In this section we describe some of these tools and processes we have built that become the basis for the research described in this paper.

- **ID Score** – This scores an event for the likelihood it is an identity fraud attempt. The events scored include account applications, new check orders, changes of address or other PII, and online payments. We examine the PII (SSN, name, address, phone, date of birth, email) for inconsistencies and anomalous associations.

- **ID Resolution** – Examines an asserted set of PII and determines the unique person behind the assertion and provide a likelihood confidence measure.

- **Identity Manipulation** – Examines the multitude of variations of asserted PII at the person level, and quantitatively assesses the amount of intentional manipulations around SSN, date of birth, name and address. For each of 317 million individuals in the U.S. we have calculated an Identity Manipulation Score that quantifies such intentional and improper manipulation.

- **Consumer Notification Service ("Not Me™")** – This is a real time alerting service that we provide to the largest consumer identity theft protection subscription services (such as LifeLock). When the system detects the use of an enrolled consumer's PII, it sends a real time alert that allows the consumer to respond "that's not me." If the consumer confirms his identity was used without his permission, the service shuts down the transaction (account application, change of address, online payment…).

These are some of our mainline tools that allow us to examine in great detail the flow of events and to perform detailed analysis and research into the dynamics of identity fraud, as described further in this paper.

Uncovering Identity Manipulation

We have identified and defined a mode of identity fraud that we call identity manipulation. Here the fraudster makes slight variations to their own PII on applications for products or services with the intention of sufficiently confusing the system, so that the application will not be matched to the fraudster's true identity. This is done to avoid linking the submitted application to either past bad activity or planned future bad activity.

Using our ID Resolution capability, we assign to every one of the billions of events and identity occurrences in our ID Network a unique person label that identifies the specific person behind that event or PII occurrence. We then sort by this unique person label into 317 million unique people that we have seen in the U.S. We note that tens of millions of the unique identities are deceased, and some are likely illegal aliens. Most of our visibility is into the credit active population (including cell phones), so our visibility is not significant in the under age 16 population.

Once the process of assigning unique person labels is completed, we now have the ability to examine the explicit variation of PII around each of these 317 million individuals. We can see much variation that falls into several categories, in particular:

- ***Normal, expected variations:*** first name nickname variations (John, Johnny, Jon, Jack…); multiple last names (very common in our society where many women take their husband's last name); use of initials for first or middle names; first/middle/last name misparsing.

- ***Typos:*** single or infrequent instances of one or a few wrong characters in any PII, particularly SSN, date of birth, address, phone, or name fields.

- ***Improper deliberate variations:*** instances of frequent or systematic variations in SSN or date of birth; more than one first name excluding nicknames; more than 4 or 5 last names; frequent suspicious subtle address variations.

We have constructed an algorithm to identify and quantify the extent of improper and deliberate PII variation that ignores the many occurrences of these normal variations and typos. We have assigned a numerical score around these variations called an Identity Manipulation Score and have scored the entire 317 million people we see in the U.S to determine the extent to which they deliberately and improperly manipulate their identity. We can examine this list from the worst offenders downward, and have found that about 8 percent of the U.S. population (more than 20 million people) appears to have engaged in such improper manipulation.

Table 2 below shows the PII variation for one such identity manipulator. We have modified the personal information in all examples to protect the privacy of the individuals, however, the important interrelationships in the variations have been preserved. In this and the following table the data is presented in columns showing the different variations in each of the column fields. The data across the rows is not coincident across the rows, so the field interrelationships are not shown in these tables.

Table 2. Example of an identity manipulator using repeated, systematic PII variations. The data is in column format, and the data across the rows are not coincident.

First Name	Last Name	SSNs	DOBs	Address
ANITRA	JOHNSON	815015642	1/11/1980	307 GRANADA DR
LATASHA	MCWILLAN	815115642	2/21/1980	1401 WALKER AVE # 331
MCCLELAND	MCWILLIAMS	815215642		1095 SPRING MEADOW DR
	MCMILLAN			303 EAST SHR
	MCNULTY			1002 5TH ST
				317 WINONA ST

In this example, the identity manipulator is using three variations of an SSN, each of which was used more than once. Similarly, we see two different dates of birth, both seen multiple times. We see systematic variation on both the SSNs and the dates of birth. We see the use of three fundamentally different first names, one of which might be a data error because it looks more like it was meant to be a last name. There are 5 different last names, with both slight and very

different variations. We see nothing unusual around the different addresses. A second example of a more severe identity manipulator is shown in Table 3.

Table 3. Example of a severe identity manipulator using repeated, systematic PII variations.

First Name	Last Name	SSNs	DOBs	Address
IRENE	ALMONE	580530044	1/7/1968	3310 ALGONQUIN AVE
LAQUINTA	CALHONE	586530044	12/21/1969	400 SCRUB OAK CT
LAQUITA	CALHOON	589489998	12/27/1969	4828 TERRACE TRL
LAQUITE	THOMPSON	589499935	1/7/1969	2600 PARK BLVD
LAQUTA	TOMSON	589539044	1/16/1969	4600 FAIR PARK BLVD
LEQUITA		590030040	1/17/1969	PO BOX 60011
QUITA		590490035	1/20/1969	2129 KINGSDALE DR
RENEE		590499937	1/27/1969	3211 MARYANN DR
RICHARD		590499938	1/27/1970	3229 KNOX ST
		590529641	1/27/1979	3424 FALCON DR
		590529941	1/27/1980	34321 FALCON DR
		590529994		3628 GLEN PARK CIR
		590530014		4709 LEONARD ST
		590530035		4719 LEONARD ST
		590530036		5161 DORMAN ST
		590530037		5163 DORMAN ST
		590530040		PO BOX 15840
		590530081		3008 GALEMEADOW DR
		590530244		1313 GLASGOW RD
		590538044		3052 BIRDSONG DR
		590929664		7063 MEADOWS DR
		590930043		7800 HILL DR TRLR 168
		590960044		4412 KEETER DR
		590980044		64 FOREST GLN

In this example we see many variations of first and last names, first names of clearly different genders, 24 different SSNs used, 11 different dates of birth. We see systematic variation around both SSN and dates of birth. We also see some examples of likely address manipulation, specifically the variations around Falcon Dr, Leonard and Dorman St. In these instances, the physical mail is likely delivered correctly by the local postman, but address matching processes don't link the addresses.

These are two specific examples of identity manipulators and demonstrate the visibility we have and examples of the kind of PII manipulation that is measured with our Identity Manipulation Score. The following table gives a summarized set of statistics for some of the worst identity manipulators we find in the U.S.

Table 4. Summary statistics for some of the worst identity manipulators in the U.S.

City	First Name	# SSNs	# DOBs	# First Names	M/F?	# Last Names	City	First Name	# SSNs	# DOBs	# First Names	M/F?	# Last Names
NY	Frank	146	7	7	n	5	Houston	Mary	32	10	5	y	9
Cleveland	Jamal	106	12	6	n	5	Phoenix	Lisa	29	21	3	n	2
St. Louis	Paula	101	7	5	y	9	Phoenix	Robert	40	3	3	n	5
NY	William	100	4	3	n	6	Phoenix	Michael	24	10	9	y	8
Miami	William	100	2	3	y	6	Atlanta	Corey	39	5	4	y	3
NY	William	69	14	6	n	13	DC	Smithton	33	5	5	n	7
Detroit	Linda	46	25	5	y	10	NY	Brent	21	12	7	n	9
San Francisco	Augustina	27	18	17	y	10	Seattle	Raymond	28	13	3	n	3
Detroit	Theresa	22	21	14	y	14	Tampa	William	34	4	3	n	4
Minneapolis	Heifi	33	28	4	n	4	Tampa	Melissa	32	7	3	n	3
Minneapolis	Joseph	48	9	5	n	6	Durham	Dorian	30	8	5	n	2
DC	Anthony	44	7	3	n	5	Miami	William	34	4	3	n	4
Seattle	Dorothy	41	11	3	n	3	Atlanta	Jody	21	2	17	y	5
Minneapolis	Alton	44	5	7	y	2	Atlanta	Dawn	24	15	3	y	3
Miami	Trisco	43	4	8	n	3	Durham	Joseph	22	9	8	y	5

In this table, the number of improper variations in the important PII categories is displayed. The "M/F?" column shows whether or not the individual has used both male and female first names. In the first row we see, for example, a "Frank" in New York who has used 146 different SSNs, seven different dates of birth, seven different first names beyond nicknames, and five different last names.

Next we show a list of the large metropolitan areas in the U.S. with the highest incidences of identity manipulation. To create this list, we group the scored population into the 1,000 three-digit ZIP code regions, compute the average Identity Manipulation Score for each region, and then rank order the regions by badness. The following table shows the results, showing which large areas in the U.S. have the worst overall occurrence of identity manipulation per capita.

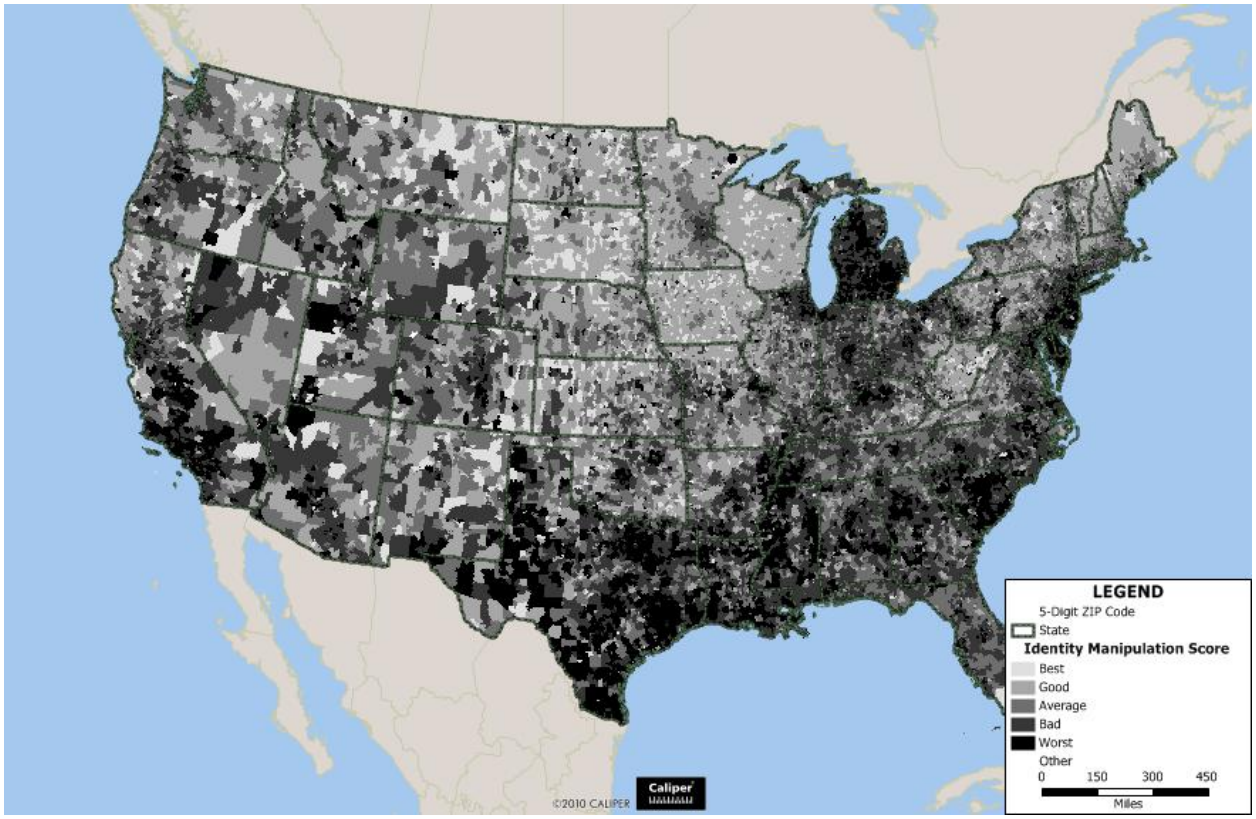
Table 5. U.S. metropolitan region (3-digit zip code) with the worst per capita identity manipulation.

Rank	3-Digit Zip	Region	Rank	3-Digit Zip	Region
1	777	Beaumont, TX	21	750	N Dallas, TX
2	799	El Paso, TX	22	197	Wilmington, DE
3	482	Detroit, MI	23	114	Queens, NY
4	485	Flint, MI	24	780	W San Antonio, TX
5	392	Jackson, MS	25	701	New Orleans, LA
6	464	Gary, IN	26	77	NE of Trenton, NJ
7	489	Lansing, MI	27	782	San Antonio, TX
8	774	SE Houston, TX	28	191	Philadelphia, PA
9	700	New Orleans, LA	29	110	E Queens, NY
10	770	Houston, TX	30	103	Staten Island, NY
11	785	Mcallen, TX	31	783	W Corpus Christi, TX
12	784	Corpus Christi, TX	32	199	S Wilmington, DE
13	751	SE Dallas, TX	33	928	Anaheim, CA (S of LA)
14	387	Greenville, MS	34	773	N of Houston, TX
15	483	NW Detroit, MI	35	711	Shreveport, LA
16	775	SE Houston, TX	36	651	S Jefferson City, MO
17	797	Midland, TX	37	926	Costa Mesa/Irvine, CA
18	48	Rural, SE Maine	38	219	SW of Wilmington, DE
19	776	Beaumont, TX	39	381	Memphis, TN
20	788	W San Antonio, TX	40	723	West Memphis, AR

With the entire table of three-digits ZIP code areas we can produce a map of the U.S. that shows the regions with the highest amount of identity manipulation. This is shown in the following figure. We see several interesting features in this map, such as areas of high activity in southern Texas and the nearby Gulf States, the East Coast, and a belt of high activity through the Ohio River Valley from Memphis through Cincinnati. Michigan sticks out like a literal sore thumb, along with the area from Louisville through Chicago. Southern California is bad, particularly in the Los Angeles region. The Midwest, in particular the upper plains states, has relatively low identity manipulation activity. Georgia also looks bad throughout, as are the regions around DC, Baltimore, Philadelphia through New York City. Upstate New York looks benign. The southern border of the U.S. is generally bad from Brownsville, TX through El Paso.

In this map and all of our geographic analysis, we have normalized to the population density so the rate of activity is correctly measured on a per capita basis. There is no *a priori* reason why higher population density areas should have a higher amount of identity manipulation, but we do see that this is indeed the case. We do generally see a correlation between the activity level of identity manipulation and the local population density, with the higher population areas generally having higher identity manipulation rates.

Figure 1. 3-digit zip code areas with the highest per capita amounts of identity manipulation. The darker the area the higher the amount of manipulation.



Dead People Applying

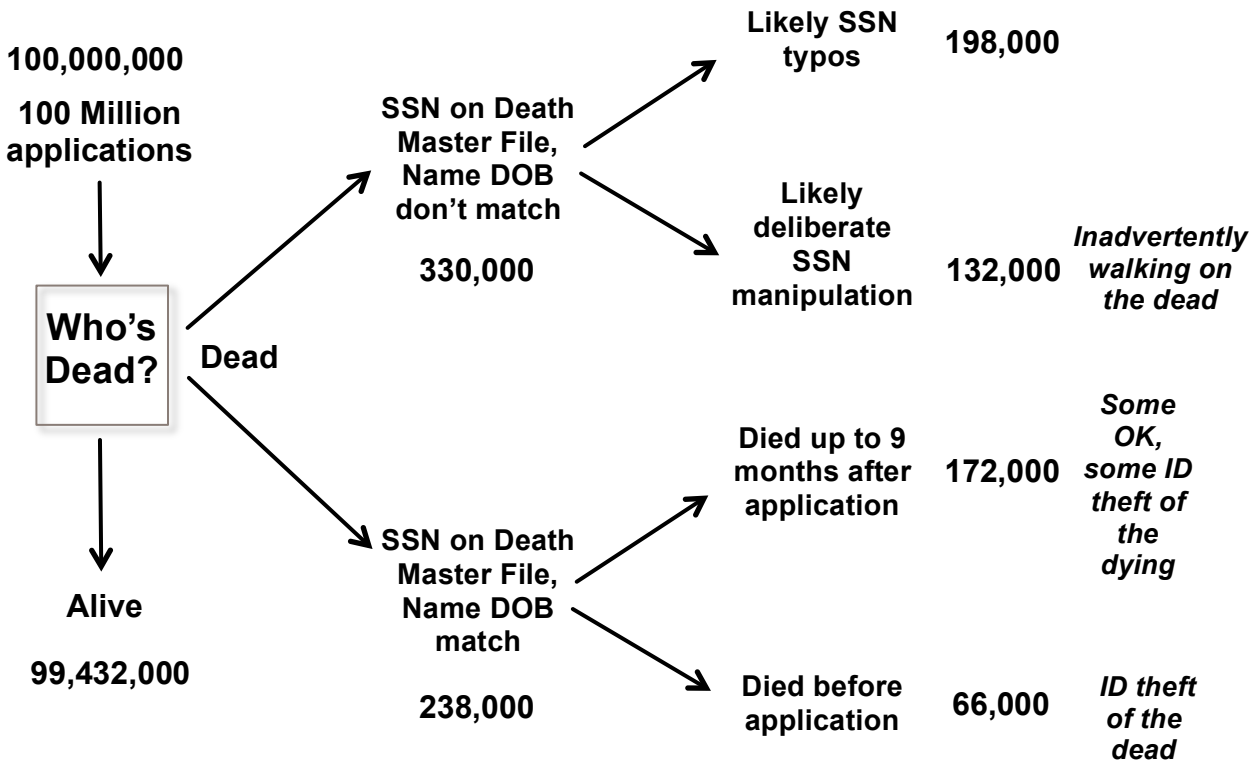
The final topic we present in this paper is analysis around the use of identities of the deceased to apply for credit products and services. Again, we are examining the instances of applications for products and services where one is required to provide correct PII, and a signature is given for permission to pull a credit report. It is illegal to misrepresent your identity in these events.

We examined 100 million applications for these products (primarily credit cards and cell phones) during the first 3 months of 2011. We compared the name, date of birth and SSNs on these applications to the data on the Social Security Administration Death Master File to find which applications used PII associated with deceased individuals and when. Of the 100 million examined, we found 99,432,000 do not match the SSA Death Master File and the remaining 568,000 do have some match. There are various ways there could be a match on the Death Master File. 330,000 of the matches showed that the SSN was on the Death Master File, but the name_date_of_birth did not match between the application and the Death Master File. Of these that had this partial match, we estimate that 60 percent (198,000) are typos and 40 percent (132,000) are deliberate SSN manipulations as described in the previous section. These

estimated 132,000 instances of SSN matches are cases where an identity manipulator inadvertently used the SSN of a deceased person.

The other group of matched applications are the 568,000 – 330,000 = 238,000 applications where the SSN, name and date of birth all matched to the SSA Death Master File. Of these, we found that 172,000 had died in the months after the application, so they were not dead at the time of application. Some unknown number of these are likely to be the misuse of the identities of the dying, but we have no visibility into how many this might be. The remaining 66,000 of these complete match applications showed that the person had died before the application, so these are clear instances of a deceased person's identity being misused. The following figure shows a summary of these categories.

Figure 2. Analysis of deceased identities being used for application for products and services. Shown is the categorization of 100 million applications, about 3 months of data into our ID Network.



This analysis was performed on the application volume seen during a three-month period, and ID Analytics sees roughly one-third of all U.S volume. When we scale to all the U.S. volume over a one-year period, we estimate that there are about 800,000 applications using deceased people's identities each year.

Summary

In this paper, we first presented a definition and description of the different modes of identity fraud and describe the nature of the problem and methods of detection. We presented a description of the data that is seen in our ID Network that provides the basis for our analyses.

There are several tools that we use for finding and quantifying different modes of identity fraud, each with its specific purpose. Using our ID Resolution and identity manipulation algorithms, we are able to find several tens of millions of consumers who have engaged in substantial and improper identity manipulation as they applied for consumer products and services.

We showed several examples of identity manipulators and described in detail how they made improper and deliberate variations in their PII for these product applications. A set of 30 examples of severe identity manipulators was shown in Table 4. We then examined the geographic distribution of where identity manipulation occurs the most.

The last analysis presented was examining the use of deceased people's identities being misused in applications for commercial products, and we found just under a million take place each year.

The data flow into the ID Network gives us tremendous and unprecedented visibility into the nature and dynamics around identity fraud and allows us to perform many special and detailed analyses to provide deep understanding into the nature of identity fraud. We use this understanding to continue to field and improve tools for use in business decision processes to find and prevent occurrences of this continuing problem.

Summary Facts Around Identity Fraud

- 20 million people have multiple SSNs associated with them. About 12 million of these are due to typos.
- 8 million people deliberately use 2 or more SSNs.
- 2 million people deliberately use 3 or more SSNs.
- 40 million SSNs have multiple people associated with them.
- More than 20 million people in the U.S deliberately manipulate their PII in an attempt to improperly obtain products and services.
- About 16 million people deliberately use multiple dates of birth.
- About 10 million spouses improperly share identity information (usually SSN).
- About 6 million parents/children improperly share identity information (usually SSN).
- At least one million children have been victims of identity theft. About 500,000 from strangers (at a rate of about 140,000 incidents per year) and about 500,000 are victimized by their parents.
- More than 2 million elderly parents are identity theft victims perpetrated by their adult children.
- About 2% of applications for credit products and services (includes credit cards, loans, cell phones…) are fraud attempts.
- About 800,000 of these applications each year are using deceased people's identities.