The University of Texas at Austin
Center for Identity

, O I BODJOHĪBOEĪ , WBMVBUJOHĪ0EFOUJUZĪ 7SJWBDZĪBOEĪ (VU I FOUJDBUJPOĪ : USFOHU I Ī CZĪ<UJMJ[JOHĪU I FĪ0EFOUJUZĪ , DPTZTUFN

2BJĪ I JI Ī* I BOH*
9B[JF I Ī5PL I CF I ĪABFFN
2]Ī: V[BOOFĪ) BSCFS

UTCID Report #1822

NOVEMBER 2018

Enhancing and Evaluating Identity Privacy and Authentication Strength by Utilizing the Identity Ecosystem

Kai Chih Chang
The University of Texas at Austin
Austin, Texas
kaichih@identity.utexas.edu

Razieh Nokhbeh Zaeem
The University of Texas at Austin
Austin, Texas
razieh@identity.utexas.edu

K. Suzanne Barber
The University of Texas at Austin
Austin, Texas
sbarber@identity.utexas.edu

ABSTRACT

This paper presents a novel research model of identity and the use of this model to answer some interesting research questions. Information travels in the cyber world, not only bringing us convenience and prosperity but also jeopardy. Protecting this information has been a commonly discussed issue in recent years. One type of this information is Personally Identifiable Information (PII), often used to perform personal authentication. People often give PIIs to organizations, e.g., when applying for a new job or filling out a new application on a website. While the use of such PII might be necessary for authentication, giving PII increases the risk of its exposure to criminals. We introduce two innovative approaches based on our model of identity to help evaluate and find an optimal set of PIIs that satisfy authentication purposes but minimize risk of exposure. Our model paves the way for more informed selection of PIIs by organizations that collect them as well as by users who offer PIIs to these organizations.

KEYWORDS

Identity, Internet of Things, Privacy, Authentication

1 INTRODUCTION

Identity authentication has been an integral part of network security to ensure and improve the security of services in the cyber world. Authentication, however, often requires the collection of personally identifiable information (PII), which increases the risk of PII exposure to identity theft and fraud criminals.

With the concept of "Smart City" growing rapidly, the relationship between people's PII and devices has become blissfully tight [4]. It is not an uncommon scene that our personally identity attributes are being collected dynamically and transmitted in today's Internet of Things society. Medical and sport devices are collecting our body temperature and heart rate. Vehicles and footbridges are recording our GPS history. Identity attributes are being collected anytime and anywhere. The diversification of PII collection, however, has opened new gateways for identity fraudsters.

In 2017, the number of identity fraud victims increased by 8% rising to 16.7 million U.S. consumers. Fraudsters stole from 1.3 million more victims in 2017 stealing $16.8 billion from U.S. consumers [8]. More comprehensive and efficacious methods and analyses are needed in order to prevent identity attributes from being compromised. Breaches often occur in unexpected places. Identity information travels in the cyber world through the Internet and eventually, a person's PII could flow to someone's devices and end up at an organization. A security incident at that organization may expose personal information that belongs to a large number of people and result in monetary loss [4]. Analyzing the relationships between people, devices, and organizations—through which PII flows—is fundamental in order to prevent fraudulent activities and enhance privacy in the Internet of Things society.

In addition to analyzing the relationships between people, devices, and organizations to minimize risk of exposure, we discuss methods to tailor the set of PIIs an organization collects to minimize the risk of exposure but at the same time achieve the same level of authentication strength. People often give personal information to organizations, for example through online forms. The purpose of providing identity attributes can vary, but is often authentication. The research question we seek to answer is that "Can we find a set of identity attributes that minimizes the risk of exposure but at the same time maximizes the authentication strength?"

To answer this question, we provide two novel approaches. The first approach is static, i.e., it only considers static relationships between PIIs. The UT CID Identity Ecosystem [4, 13] developed at the Center for Identity (CID) at the University of Texas (UT) at Austin has constructed a graph-based model of people, devices, and organizations. It provides a framework for understanding the value, risk and mutual relationships of identifiable information attributes. Every attribute is modeled as a graph node which has several properties, while the relationships between identity attributes are modeled as edges. We take two of node properties, uniqueness and risk, into account in our first approach. The next approach is a dynamic method. The Ecosystem tool, in its present form, is capable of using Bayesian inference to perform three chief kinds of analyses: 1) analyzing the risk of exposure, 2) inferring the most likely source of a breach, and 3) calculating the expected cost of attributes. We are utilizing the underlying formulae of these analyses to answer our question in this research. Finally, we perform experiments and discuss the general identity model in Ecosystem and the identity set required in specific user cases. The remainder of this article is structured as follows. Section 2 presents the related work of privacy, authentication, and identity in the Internet of Things (IoT). Section 3 provides a brief description of our methodology. Section 4 includes a comprehensive analysis for evaluation. Section 5 concludes our research and gives insights for future works.

2 RELATED WORK

2.1 Privacy

Security and privacy issues are major obstacles to the implementation of the Internet of Things. Some researchers have addressed these issues and suggested various security and privacy measures [10][1]. Weber [11] has mentioned that the main challenge for privacy in the context of IoT remains the management of the vast

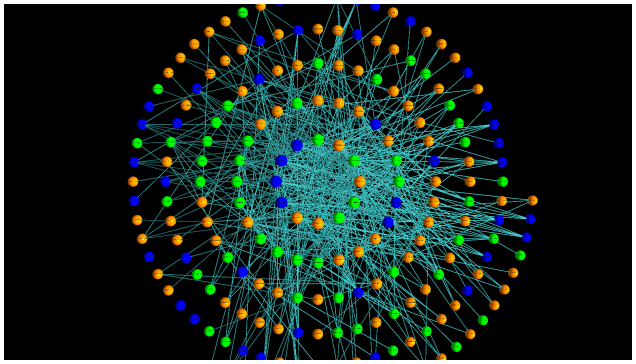

Figure 1: A snapshot showing the identity attribute graph in the Ecosystem.

amount of data collected. Concerns are raised over access of personally identifiable information (PII) pertaining to IoT devices and organizations. It will be something entirely different for a connected set of organizations and machines to have access to and to utilize information about the environment in which one behaves and exists [12].

2.2 Authentication

There is much research done in the area of securing IoT. Identity authentication has been considered one of the main issues in the IoT world. Mahalle et al. [6] presented an efficient and secure integrated authentication and access control protocol. They also presented a mutual authentication protocol that they integrated with novel and secure approaches for access control in IoT.

Most of the IoT products available in the market are incapable of securing identities and hence lead to various identity breaches. Jan et al. [5] proposed a lightweight secure authentication algorithm that verifies the identities of the clients and servers participating in the network. However, no secure solution in the world can combat all types of attacks. That is one reason why our research aims to reduce the probability of being attacked by tailoring the set of PIIs one gives away.

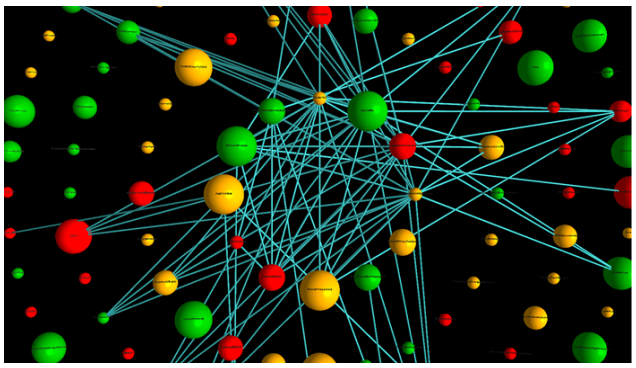

Figure 2: A snapshot showing identity attributes with utilization of property filters.

The existing security solutions mainly provide security approaches for a general IoT, and there is little authentication scheme particularly designed for the U2IoT architecture (i.e., unit IoT and ubiquitous IoT). In 2015, Ning et al. [7] proposed an aggregated-proof based hierarchical authentication scheme for the U2IoT architecture, establishing trust relationships via the lightweight mechanisms, and applying dynamically hashed values to achieve session freshness.

3 METHODOLOGY

In this section, we seek to answer the question "For a given number of identity attributes, what is the optimal set of PII to maximize authentication strength while minimizing the risk of exposure?" We came up with two different solutions based on how we extract required information from the Ecosystem probabilistic model and how we interpret that information.

3.1 The Identity Ecosystem

Before we start going through our approaches, we provide a high level introduction to our Identity Ecosystem first. We have designed and implemented the Identity Ecosystem at the Center for Identity at the University of Texas at Austin as a valuable tool that models identity relationships, analyzes identity frauds and breaches, and answers several questions about identity management. It stores known data about identity attributes in a probabilistic model, and performs Bayesian Network-based inference to calculate the posterior effects on each attribute. It presents identity attributes as nodes and various types of connections between nodes as edges. Each node has its own properties such as type of node, risk of exposure, and intrinsic monetary value. Figure 1 shows the typical set of identities for people, devices, and organizations. Nodes for people are colored in orange, nodes for devices are colored in blue, and nodes for organizations are colored in green. The Ecosystem Graphical User Interface (GUI) can color and size attribute nodes based on various properties. In Figure 2 nodes are colored based on their risk of exposure and are sized based on their value.

So far the Ecosystem is capable of answering three questions relevant to the overall risk and liability of any person in terms of managing identity attributes. We do not model a specific person's identity graph because we want to have a more generic and comprehensive analysis for the universal relationships of identities and miscellaneous risks for identity management. The first question Ecosystem can answer is "When a set of attributes is exposed, how does it affect the risk of other attributes being exposed?" For instance, if the SSN of an individual is compromised, what are the most risky node items that fraudsters might try to obtain after that? Multiple attributes can be selected as evidence (i.e., exposed PII) at the same time. It also shows potential loss after such a breach. The next question Ecosystem can answer is "If a set of attribute have been exposed, what was the most likely origin of the breach?" If an individual finds out that his or her health insurance records have been compromised, the Ecosystem can help to detect the most probable origin of breach. The last question Ecosystem can answer is "What is the total cost of an attribute being exposed?" We would like to find out how an attribute's exposure increases the risk of other attributes' getting exposed, so that an attribute incurs not only its own intrinsic cost, but also some expected costs downstream.

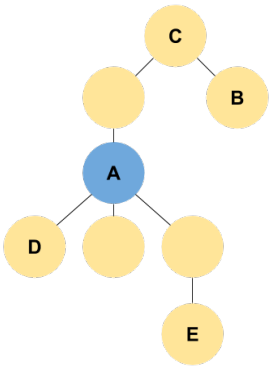

Figure 3: An illustration for ancestors and descendants.

We represent the Identity Ecosystem as a graph $G(V, E)$ consisting of N attributes $A_1, ..., A_N$ and a set of directed edges as a tuple $e_{ij} =< i, j >$ where A_i is the originating node and A_j is the target node such that $1 \leq i, j \leq N$. Each edge e_{ij} represents a possible path by which A_j can be breached given that A_i is breached.

3.2 Static Approach

Our first solution is a static approach. We have defined a person's identity as a set of informational data that are linked to a person. Each such piece of information is called an attribute. Attributes can be classified in many ways depending on their properties, such as whether or not an attribute is widely used, how accurately it can be verified, etc. Among several different properties for attributes we have defined, we are using Risk and Uniqueness for our first solution. We identified the definition and the measurement of Risk and Uniqueness as follows:

- Risk (shows the risk of exposure): Low, Medium, High.
- Uniqueness (shows how unique the PII is for the individuals who have it): Individual, Small Group, Large Group.

The uniqueness of an identity determines the strength of the identity [3]. By referring to the Bayesian Network Model in the Ecosystem, we could obtain the uniqueness of each PII attributes. A strong identifier uniquely identifies an individual in a population, whereas a weak identifier can be applied to many individuals in a population [2]. Thus, we assign a score to each attribute based on their level of uniqueness. The stronger its uniqueness is, the higher score it obtains. We assign 2 points to attributes with uniqueness level of "Individual". We assign 1 point to "Small Group" and 0 point to "Large Group". To minimize exposure, we tend to choose attributes with lower risk. Accordingly, we assign 2 points to attributes with risk level of "Low", 1 point to level of "Medium", and 0 point to level of "High".

A total score of the combination of risk and uniqueness can be computed as

$$S = \alpha U + \beta E$$

, where U is the score of uniqueness and E is the score of risk. Given n identity attributes $A_1, ..., A_n$, let S_i be the total score of A_i such that $1 \leq i \leq n$. Sort these n attributes according to S in descending order. The sorted list is what we want.

3.3 Dynamic Approach

Our second solution is a dynamic approach. Different from the first approach where we utilized identity attributes' intrinsic properties, we are using Bayesian inference to perform our second solution. Each attribute A_i is labeled with a Boolean random variable, denoted $D(A_i)$, which is *true* if the attribute has been exposed/breached and $false$ otherwise. Each attribute has a prior probability $P(A_i)$ of it getting exposed on its own. Given a target node, by combining these two variables, we try to calculate the accessibility of each ancestor point to the target point and the influence of each target point for each descendant point. Finally, combine these two data to sort all the points for analysis. Next we briefly talk about our calculation.

First we take ancestors into account. In Figure 3, we can see that node C is node A's ancestor, but node B is not. Every ancestor of node A has a path that can lead to node A. Let, $ANCESTORS(A_i)$ be the set of ancestors of A_i. For every attribute A_i, if $A_k \in ANCESTORS(A_i)$ is exposed, we want to know the posterior probabilities for A_i. We set the exposure evidence values $D(A_k)$ for each node in the set $ANCESTORS(A_i)$ to *true*, and use Bayesian inference to compute the posterior probabilities $P'(A_i)$. Now, given the $P'(A_i)$ values, it is easy to compute the percentage increase in the risk as $(P'(A_i) - P(A_i))$ given that $D(A_k) = true$. Hence, the sum of the percentage increase of A_i can be computed as $C = \sum_k (P'(A_i) - P(A_i))$ and we call it the "Accessibility". We sort these N attributes according to their accessibility in ascending order. The low value of accessibility of one attribute indicates that it is more difficult to get to this attribute than others. One reason of its low value could be the small size of the set of its ancestors, which makes it harder to get to this attribute due to only few entrances. It also makes it more difficult in the process of getting this attribute and hence makes it have stronger strength of authentication. After sorting, we give

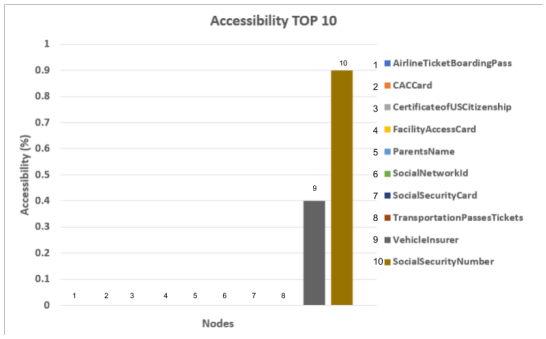

Figure 4: Top 10 nodes with the lowest accessibility.

Table 1: List of top 20 nodes.

Top 20 attributes with static approach			
AutomobileLocator	BankAccount	BloodSample	BloodType
CertificateUSCitizenship	EmailAccount	MedicaidCard	MedicalHistory
MedicareCard	MilitaryId	MilitaryServiceRecord	OtherPassport
PrescriptionNumber	ProfessionalRegNum	PublicAssistanceCards	SocialNetworkAccount
USPassport	UtilityAccounts	VehicleLoanNumber	Visa

Table 2: List of 8 nodes with 0 point in static approach.

8 attributes with 0 point in static approach			
DateofBirth	Ethnicity	EyeColor	Gender
HairColor	Weight	Height	Name

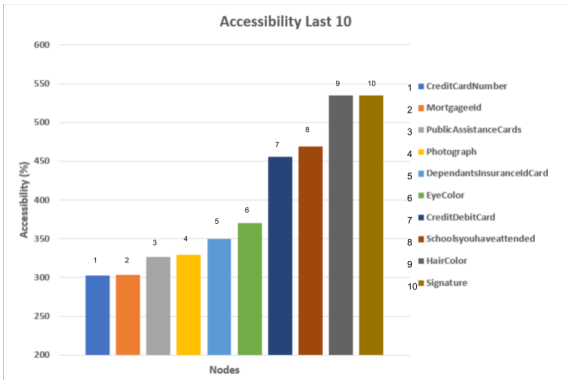

Figure 5: Last 10 nodes with the highest accessibility.

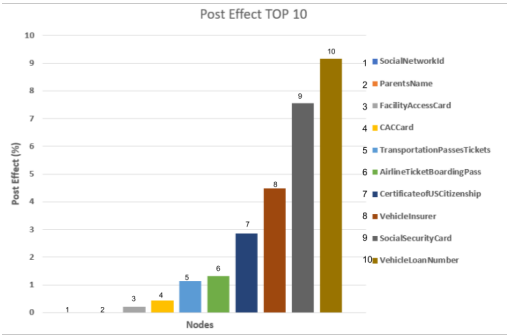

Figure 6: Top 10 nodes with the lowest post effect.

each attribute a score. The former the order of the attribute is, the higher the score it get.

Next, we consider the descendants. In Figure 3, both node D and node E are node A's descendants. So anything that happens to node A would also has some influence on node D and node E. Let, $DESCENDANTS(A_i)$ be the set of descendants of A_i and attribute in $DESCENDANTS(A_i)$ be A_k. We set the exposure evidence values $D(A_i)$ for the nodes in the given set to $true$. We compute the posterior probabilities $P'(A_k)$ for each attribute $A_k \in DESCENDANTS(A_i)$. Now, given the $P'(A_k)$ values, it is easy to compute the percentage increase in the risk as $(P'(A_k) - P(A_k))$ when A_i is exposed. Thus, the total increase of the descendants set $DESCENDANTS(A_i)$ is $E = \sum_k (P'(A_k) - P(A_k))$ and we call it the "Post Effect". We sort these N attributes according to post effect values in ascending order. The low value of post effect indicates its low risk of exposure. Instead of only taking one attribute A_i into account, we seek to analyze how it will effect all its descendants and choose the one that minimize the impact. After sorting, we give each attribute a score. The former the order of the attribute is, the higher the score it gets.

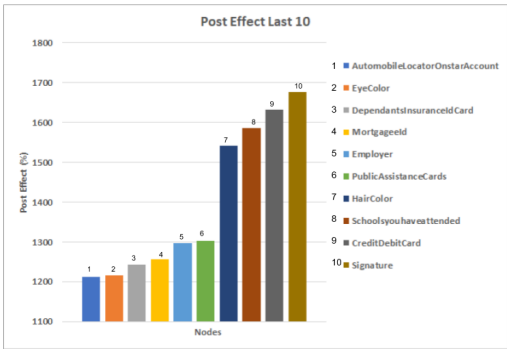

Figure 7: Last 10 nodes with the highest post effect.

The total score of accessibility and post effect can be computed as

$$S = \alpha A + \beta P$$

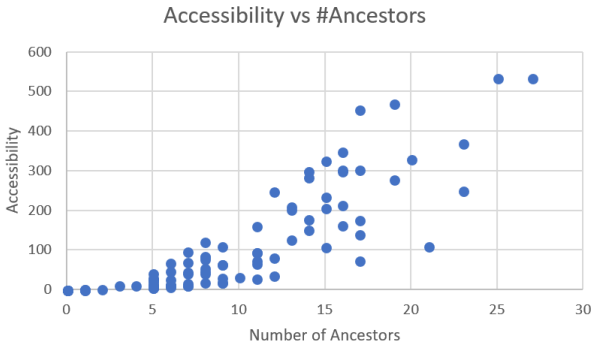

Figure 8: The scatter diagram with vertical axis as Accessibility and horizontal axis as number of ancestors.

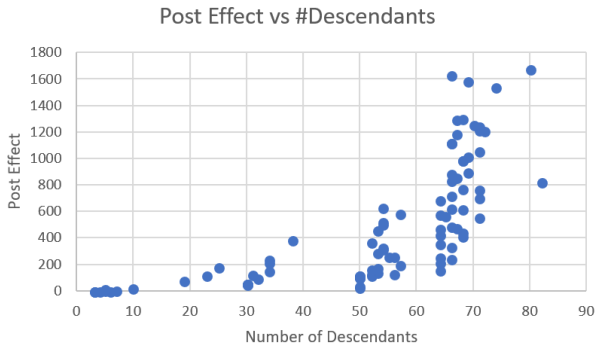

Figure 9: The scatter diagram with vertical axis as Post Effect and horizontal axis as number of descendants.

, where A is the score of accessibility and P is the score of post effect. Sort these n attributes according to S in descending order. The sorted list of attributes is what we want.

4 EMPIRICAL STUDY

In this section, we introduce how we get and extract the resource data. Then we show our results and discussion with graphics and tables coming together.

4.1 ITAP Data

The Identity Ecosystem takes ITAP's output as its input. Identity Threat Assessment and Prediction (ITAP) is a risk assessment tool that increases fundamental understanding of identity theft processes and patterns of threats and vulnerabilities. ITAP captures and models instances of identity crime from a variety of sources, and then aggregates this data to analyze and describe identity vulnerabilities, the value of identity attributes, and their risk of exposure. Through the raw data collected from news stories and other sources, ITAP aims to determine the methods and resources actually used to carry out identity crimes; the vulnerabilities that were exploited; as well as the consequences of these incidents for the individual victims, for the organizations affected, and for the perpetrators themselves.

The ITAP database is a large, structured, and continually growing repository of such information, with approximately more than 5,000 incidents captured in the model to date. The cases analyzed occurred between 2000 and 2017. The version of data set we are using is 2017 version and it was extracted in May, 2018 [9].

4.2 Results

In this section, we are going to demonstrate all statistic results and we will discuss what insights these numbers give us in the next section. In order to compare our two methods, we are going to see how much information we can derive from the result of two different approaches. There are more than 500 identity nodes in the ITAP data set so far. We first applied the static approach on our ITAP data. By doing so, the maximum point should be 4 points. The mean of the score of this data set is 2.358 points which is only 58.9% of the maximum points. We have listed the top 20 nodes in table 1. Every node in this table gains 4 points which is the highest score in the data set. We have also listed nodes that gain no points after applying the static approach in table 2. These nodes have lower uniqueness and at the same time higher risk of exposure.

Not only have we studied our own data, but we also have focused on user cases. For instances, what identity attribute set is required when applying for Naturalization at U.S. Citizenship and Immigration Services (USCIS)? Required identities in N-400 form[1] are shown in table 3. The mean of the score of this data set is 1.519 points which is 37.9% of the maximum points and it is 64.4% of the whole data set score.

We also applied our dynamic approach on our ITAP data set. Here we do not show the score of each identity node but we show the actual accessibility result. Figure 4 shows the top 10 identity attributes with the lowest accessibility. These nodes are much more difficult to access in the Identity Ecosystem graph. Figure 5 shows the last 10 identity attributes with the highest accessibility. These nodes are the 10 easiest identities for fraudsters to access. The mean of Accessibility is 107.17%. Figure 6 shows the top 10 identity attributes with the lowest post effect. These nodes have the lowest impact on other attributes when they are breached. Figure 7 shows the last 10 identity attributes with the highest post effect. These nodes have the largest impact on other nodes which might cause a large amount of monetary loss. The mean of Post Effect is 471.02%. Combining these two properties together, we show the top 12 nodes that minimize the risk and maximize the authentication strength in table 4.

Same as above, we applied the dynamic approach on the USCIS naturalization data set. Here we also show the actual value of Accessibility and Post Effect rather then showing the rank score of the whole data set since we think they give more insight than the rank score can provide. The mean Accessibility is 126.08%, which is 17.65% more than the ITAP data set. The mean Post Effect is 501.49%, which is 6.4% more than the ITAP data set. All the above results are shown in Table 5.

[1] "U.S. Citizenship and Immigration Services", https://www.uscis.gov/ (accessed May, 2018).

Table 3: List of identity attributes in N-400 form.

Data set required at USCIS			
1. MilitaryId	2. MilitaryServiceRecord	3. Email	4. SSN
5. PhoneNumber	6. TravelHistory	7. Fingerprints	8. ParentsName
9. ParentsOccupation	10. SpouseInfo	11. Address	12. BirthCertificate
13. Hometown	14. School	15. Organization	16. Signature
17. Citizenship	18. ZipCode	19. Name	20. DateofBirth
21. Height	22. Weight	23. Gender	24. EyeColor
25. HairColor	26. Ethnicity	27. CrimeHistory	28. Age

Table 4: List of 12 nodes with the highest score.

Top 12 attributes result from the dynamic approach			
1. CACCard	2. AirlineTicketBoardingPass	3. FacilityAccessCard	4. ParentsName
5. SocialNetworkId	6. CertificateUSCitizenship	7. TransportationPassTickets	8. SocialSecurityCard
9. LicensePlate	10. Last4digitsofSSN	11. MothersMaidenName	12. Occupation

4.3 Discussion

In the previous experiments we used static and dynamic methods to find the optimal data set so that it maximizes the privacy and authentication strength while minimizing the risk of exposure. Here the coefficients α and β are both set to 1. The static method uses the intrinsic characteristics of the data set of ITAP to distinguish between its uniqueness to determine its ability to authenticate and the basic risk to determine the risk at specific node.

In addition to answering our research questions, the results also gave us some interesting insights. As mentioned earlier, the overall average score of the data set is only 2.36, less than 60 percent of the maximum score. This indicates the incomprehensiveness of identity attributes nowadays. On the other hand, there still exists some high uniqueness and low risk nodes. The highest score is 4 points, which is two plus two. High degree of uniqueness and low risk are already the best choices for data sets that result in strong authentication capabilities. From table 1, we know that there are only 20 identity attributes that can reach full marks, and the proportion is 21% of the entire set, accounting for about one-fifth of the whole. Biometric identities like blood sample are included. This type of identity has a high degree of discrimination and authentication capabilities in the current cyber world. On the whole, static methods lead us to the conclusion that it is better to use these identity attributes for authentication.

For the dynamic method, we utilized the statistical tool Bayesian Network inference. We first use each node's ancestors to calculate each node's Accessibility. The more difficult to get to this attribute, the harder it is to obtain in the process. Figure 8 shows the distribution map of the Accessibility over the number of ancestors. The result is reasonable. The more the ancestors, the higher the possibility of a node's Accessibility. Then we calculate the Post Effect of each node. It can be seen from the distribution of Figure 9 that the number of descendants is almost proportional to the Post Effect. If a node has many paths to other nodes, its influence will also increase.

The final result of combining these two new features is not the same as the static result. The result is not so intuitive. In Table 4, the first few points have almost no edges, which made them much higher in the score of accessibility, but this point that has no incoming edges in Ecosystem graphic model is also our best node who obtains authentication ability because no path can reach them, making them the most unique in the graphical structure. On the other hand, if this kind of node also obtains few out-degree edges, or few descendants, this node is probably isolated from other clusters of nodes.

We see another interesting insights in the scatter diagrams. When we derived the Accessibility and the Post Effect, we thought that the better the number of them, the better the rank of the node is. In figure 8, close to the intersection point of 100 for accessibility and 15 for number of ancestors, there are some nodes with their accessibility close to the mean of 107% (±10%) with the number of ancestors between 13 and 20. We look between 13 and 20 because it is close to the half value of the maximum number of ancestors. Nodes that possess high ranking in static method such as StudentLoanNumber, MilitaryId, BloodSample, etc are in this area. Using the same point of view to observe Figure 9, we can derive some similar results. The mean Post Effect is 471%. Taking the range of ±10% and the number of descendants 40 to 50 into account, nodes with high ranking in static methods like MedicareCard, MilitaryServiceRecord, StudentLoanNumber, etc are discovered in this area. If only target on the quantity of the Accessibility and the Post Effect, the high-ranking nodes should be the in the left-bottom area of the diagram, but in fact they are around the area we just talked about. Therefore, identity attributes that retain the better Accessibility or better Post Effect would not always be the best choice. We need to take graphic model structures like the number of ancestors and in/out degree edges which we categorize them as the "physics" of identity into account. This leaves us more space to investigate and requires further research in the future.

For the USCIS data set, in static approach, the mean value is 1.52 which is only 37.96% of the maximum score. It indicates that in the intrinsic characteristic aspect, this data set needs more improvement, which means the identities required in N-400 form are not that appropriate and need to be modified. In dynamic approach, the

Table 5: Statistic results for both approaches.

Statistic results for both approaches			
	General	USCIS	Percentage (USCIS/General)
Static	2.358	1.519	64%
Accessibility (mean)	107.17%	126.08%	17%
Post Effect (mean)	471.02%	501.49%	6%

results are different. Its mean value of accessibility is 126.8% which is 17% higher than the average value. So there are more identities that people can easily obtain in process in this data set. The mean value of post effect is 471%, which is only 6% higher than the average value. So the expected loss in applying for the N-400 form is nearly the same as general loss in the society.

The current limit for static approach is essentially the degree of leveling properties. ITAP currently only allows these two characteristics to be divided into three levels so that these nodes currently cannot have a more detailed distribution of results. The limitation of the dynamic method is that it is calculated using statistical tools. However, the number of identity attributes we have is not enough. At the same time, we have not assigned enough edges between those nodes. If the number of nodes is not enough, it will be less complete in the result.

5 CONCLUSION

In this paper, we try to find out ways to determine when given a set of PII, how to identify its authentication capabilities and risks, and simultaneously can we find an alternative or optimal set to replace it. We provide static and dynamic approaches. The static method uses the intrinsic properties of the data set of ITAP to determine its authentication strength based on its uniqueness and to determine its privacy level based on its prior risk. The dynamic method uses the statistical tool Bayesian Network model to help with analysis. For a specific node, we first find out all its ancestors. Then we make every ancestor breached to calculate the sum of the percentage increases on this node and we call this value the Accessibility. Next, we calculate the post effect of this node to see when this node is breached, and what the impact is on the overall descendants. When giving away an identity with high Post Effect, the impact on the individual person is also going to be high with respect to monetary loss. So using this kind of attributes for identity certification will reduce our privacy and safety.

We use the data set collected from ITAP for experiments. The static and dynamic methods give us different but meaningful results. Static method allows us to distinguish from the essence that it is better to use these points for authentication. The result is more intuitive. The results obtained by the dynamic method which is based on the probability tool reflect the particular role of the specific node in the structure of the Ecosystem graphical model. The lower the accessibility, the lower number of its in-degree edges, concurrently the number of ancestors may still be many. The lower the Post effect, the lower the number of out-degree edges, meanwhile the number of descendants may still be many. We also use the identity set required in the N-400 form used for naturalization on the USCIS website to determine the risk of this data set and the effectiveness of the validation based its overall score.

As the ITAP project continues collecting data, theories and technologies developed in or from this research can be customized along the way to minimize our identities' risk of exposure and maximize the privacy and authentication strength in nowadays Internet of Things society.

REFERENCES

[1] L. Atzori, A. Iera, and G. Morabito. The internet of things: A survey. *Computer Networks*, 54(15):2787 – 2805, 2010.

[2] E. Bertino, F. Paci, and N. Shang. Keynote 2: Digital identity protection - concepts and issues. In *2009 International Conference on Availability, Reliability and Security*, pages 69–78, March 2009.

[3] Y. Cao and L. Yang. A survey of identity management technology. In *2010 IEEE International Conference on Information Theory and Information Security*, pages 287–293, Dec 2010.

[4] K. C. Chang, R. N. Zaeem, and K. S. Barber. Internet of things: Securing the identity by analyzing ecosystem models of devices and organizations. In *2018 Association for the Advancement of Artificial Intelligence Spring Symposium*, March 2018. To Appear.

[5] M. A. Jan, P. Nanda, X. He, Z. Tan, and R. P. Liu. A robust authentication scheme for observing resources in the internet of things environment. In *2014 IEEE 13th International Conference on Trust, Security and Privacy in Computing and Communications*, pages 205–211, Sept 2014.

[6] P. N. Mahalle, B. Anggorojati, N. R. Prasad, and R. Prasad. Identity establishment and capability based access control (iecac) scheme for internet of things. In *The 15th International Symposium on Wireless Personal Multimedia Communications*, pages 187–191, Sept 2012.

[7] H. Ning, H. Liu, and L. T. Yang. Aggregated-proof based hierarchical authentication scheme for the internet of things. *IEEE Transactions on Parallel and Distributed Systems*, 26(3):657–667, March 2015.

[8] A. Pascual, K. Marchini, and S. Miller. 2018 identity fraud: Fraud enters a new era of complexity. Technical report, Retrieved from Javelin Strategy & Research: https://www.javelinstrategy.com/coverage-area/2018-identity-fraud-fraud-enters-new-era-complexity, 2018.

[9] Center for Identity. Itap data, 2017. Unpublished raw data.

[10] S. Sicari, A. Rizzardi, L. Grieco, and A. Coen-Porisini. Security, privacy and trust in internet of things: The road ahead. *Computer Networks*, 76:146 – 164, 2015.

[11] R. H. Weber. Internet of things: Privacy issues revisited. *Computer Law & Security Review*, 31(5):618 – 627, 2015.

[12] B. D. Weinberg, G. R. Milne, Y. G. Andonova, and F. M. Hajjat. Internet of things: Convenience vs. privacy and secrecy. *Business Horizons*, 58(6):615 – 624, 2015. Special Issue: The Magic of Secrets.

[13] R. N. Zaeem, S. Budalakoti, K. S. Barber, M. Rasheed, and C. Bajaj. Predicting and explaining identity risk, exposure and cost using the ecosystem of identity attributes. In *Security Technology (ICCST), 2016 IEEE International Carnahan Conference on*, pages 1–8. IEEE, 2016.

: QPOTPSFEĪ) Z


HID


gemalto
security to be free


GENERALI
GLOBAL ASSISTANCE
IDENTITY PROTECTION SERVICES


IDEMIA
augmented identity


LifeLock
Guarantee Your Good Name


TransUnion

WWW.IDENTITY.UTEXAS.EDU