The Thesis Committee for Chia-Ju Chen
certifies that this is the approved version of the following Thesis:

Statistical Analysis of Identity Risk of Exposure and Cost Using the Ecosystem of Identity Attributes

APPROVED BY

SUPERVISING COMMITTEE:

K. Suzanne Barber, Supervisor

Razieh Nokhbeh Zaeem

Statistical Analysis of Identity Risk of Exposure and Cost Using the Ecosystem of Identity Attributes

by

Chia-Ju Chen

THESIS

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE IN ENGINEERING

THE UNIVERSITY OF TEXAS AT AUSTIN

December 2019

Acknowledgments

I would like to give special thanks to my research supervisor Dr. Barber, Dr. Zaeem and my parents to support my study. Also we wish to thank the Center for Identity Partners (`https://identity.utexas.edu/strategic-partners`) for their contributions to this research effort.

Abstract

Statistical Analysis of Identity Risk of Exposure and Cost Using the Ecosystem of Identity Attributes

Chia-Ju Chen, M.S.E.
The University of Texas at Austin, 2019

Supervisor: K. Suzanne Barber

Personally Identifiable Information (PII) is often called the "currency of the Internet" as identity assets are collected, shared, sold, and used for almost every transaction on the Internet. PII is used for all types of applications from access control to credit score calculations to targeted advertising. Every market sector relies on PII to know and authenticate their customers and their employees. With so many businesses and government agencies relying on PII to make important decisions and so many people being asked to share personal data, it is critical to better understand the fundamentals of identity to protect it and responsibly use it. Previously developed comprehensive Identity Ecosystem utilizes graphs to model PII assets and their relationships and is powered by empirical data from almost 6,000 real-world identity theft and

fraud news reports to populate the UT CID Identity Ecosystem. We analyze UT CID Identity Ecosystem using graph theory and report numerous novel statistics using identity asset content, structure, value, accessibility, and impact. Our work sheds light on how identity is used and paves the way for improving identity protection.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Personally identifiable information (PII) is any data that could potentially be used to recognize a particular person, and it is commonly used in both physical and cyber spaces to perform personal authentication. Identity theft is the fraudulent acquisition and usage without permission of a person's PII. A modern authentication process usually requires collection of PII and increases the risk of exposure to identity theft and fraud criminals.

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 a total of $16.8 billion from U.S. consumers [16]. More intelligent and comprehensive approaches should be provided to thwart the crime of identity theft.

In order to model the identity ecosystem, an intuitive approach is to analyze the components from both cyber and physical aspects. Modern society seamlessly merges online and offline PII attributes. Examples of on-line attributes are one's social media accounts, on-line shopping patterns, passwords, and email accounts. Off-line attributes are those related to the physical world such as bank accounts, credit and debit cards, Social Security Number, and

one's physical characteristics.

The UT CID Identity Ecosystem developed at the Center for Identity (CID) at the University of Texas (UT) at Austin constructed a graph-based model of people, devices, and organizations [25]. It models the relation between PII as a Bayesian Network, and performs interference for possible sources of breaches and cost if the source is compromised. It provides a framework for understanding the value, risk and mutual relationships for pairs of PII attributes. Each vertex represents an attribute whereas edges in-between imply the relationship.

For data source of ecosystem, The Identity Threat Assessment and Prediction (ITAP) [27] project is leveraged. ITAP is developed to focus on gathering identity theft information from news stories, structuring this information, analyzing it, and discovering trends and characteristics.

We obtained UT CID Identity Ecosystem and the ITAP collection as the data source. Based on this graph-based network of identity, we have designed and implemented a visualization framework that facilitates understanding of the whole risk network rather than reviewing unstructured raw news feed data from ITAP. We introduce three main statistical evaluation criteria: (1) Traditional pie, bar, and scatter plots of vertex or edge specific values are employed to show the distribution. (2) Centrality measures such as degree, closeness, and betweenness centrality are utilized and hence illustrate each PII with certain structural features. (3) Strongly Connected Components (SCC) are applied to distinguish groups of PII that are interconnected. With these

criteria, our visualization framework can prototype the identity system with detailed features such as PII that are most efficient to spread the information if breached, or PII that are aggregated as a group that will easily be traversed if one member is already compromised. The main contribution of this thesis includes the application of sophisticated graph theoretical concepts to reveal unprecedented insights into PII and the relationship between PII attributes.

The remainder of this article is structured as follow. Chapter II elaborates on the importance of statistical analysis of the Ecosystem tool and the set of measurements to be included. Chapter III presents a comprehensive evaluation and takeaways from the results. Chapter IV includes the related work of the identity ecosystem, identity theft, and related government reports. Chapter V concludes the research and gives insights for future work.

Chapter 2

Statistics Based on Ecosystem

Identity is the new currency. Today, it is virtually impossible to buy anything, access government services, enter an airport, a place of work, an amusement park or a computer network without providing identity information. Identity information is valuable to the individual, to corporations, to government agencies but unfortunately to criminals as well. Privacy and security are vital to protect these valuable identity assets. To protect an asset we must know that asset. To protect identity information currency assets, we must know the complete inventory of those assets.

Leveraging the empirical ITAP data, the UT CID Identity Ecosystem project is building the canonical inventory describing different types of identity assets, the valuation of those assets and the connectedness of those assets that produces a type physics in the Identity Ecosystem. Identity assets cannot be treated as a simple "list of data" but must be managed as a complex and dependent network of assets with different entry points, vulnerabilities and ever-changing values and risk. A deeper understanding of this network of identity assets will result in a better understanding of how to protect, use, and monetize (for the legitimate reasons) these identity assets.

The Identity Ecosystem is a valuable previously implemented tool that models identity liaison, analyzes identity fraud and breaches, and answers several questions about identity risk management [25]. It maps identity attributes in a probabilistic model and performs Bayesian network-based inference to determine the posterior effects on each attribute. The Identity Ecosystem models individual identity attributes as nodes whereas edges in-between indicate various types of connections.

Each vertex includes different properties such as type of node, risk of exposure, and intrinsic monetary value. The Ecosystem Graphical User Interface (GUI) can color and size nodes based on their properties independently. Figure 2.1 shows an example snapshot in which the nodes are colored based on their risk of exposure and are sized based on their liability value.

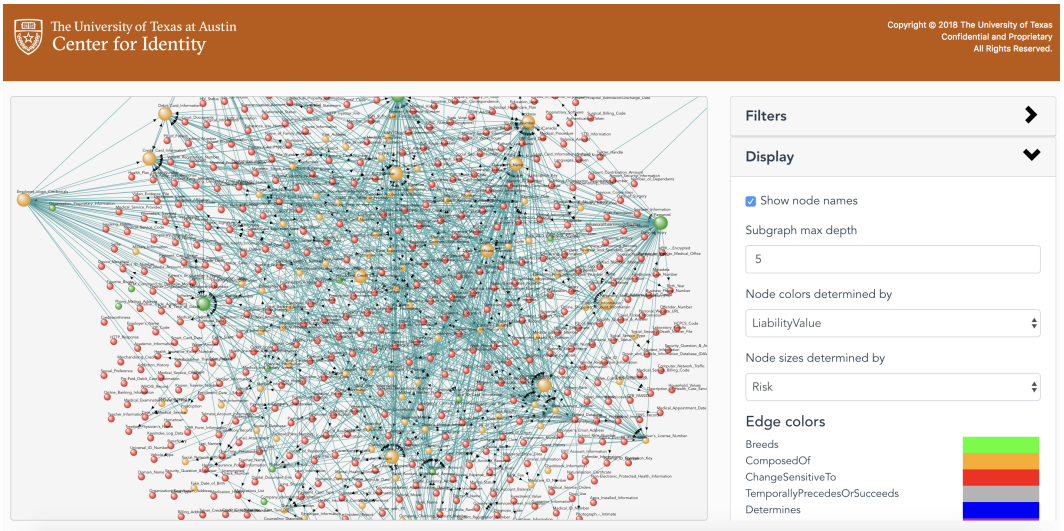

Figure 2.1: Background: A snapshot showing previously developed UT CID Ecosystem attribute graph.

The Identity Ecosystem takes ITAP's output as its input. Identity Threat Assessment and Prediction (ITAP) is a risk assessment framework that characterize the of process for identity theft and patterns of vulnerabilities [26]. ITAP captures instances of identity crime from diverse sources, further aggregates this data to portrait the identity vulnerabilities, the value of identity attributes, and their risk of exposure.

Continually collecting raw text data from various sources of news feeds stories, ITAP aims to determine the approaches and resources actually used to carry out identity crimes; the vulnerabilities that were exploited during the process; as well as the consequences according to these incidents for the individual victims, the organizations affected and the perpetrators involved. The ITAP database is a comprehensive, structured, and continually growing repository of such information, with approximately 6,000 incidents captured so far. The cases span from 2000 to 2018.

The research questions we seek to answer in this thesis are that "Given a network of identity, what are the underlying characteristics? Among the large amount of PII, what are the ones that are most likely sources for breaches?". As in the real-world, an individual might hold a variety of PII, ranging from banking information to electronic device logs to personal attributes. Some of those PII attributes are more important to protect since they are more likely to expose one's identity. For example, intuitively, a Social Security Number (SSN) or passport number are considered to be more critical PII since they map directly to an individual, and can be breached with high probability.

Another question we aim to answer is “If a certain PII is compromised, what is the cost for the fraudster to acquire all the other accessible PII?”. For a real-world analogy, if the name and email address of a person are exposed, what is the cost for the identity thieves to collect an individual’s credit card information or SSN record? Furthermore, (1) how efficient is it to acquire all the rest of the PII network? and (2) how likely is it for identity thieves to possess certain PII? According to these measures, we observe on “Given the PII that spread the information flow most quickly or serve as the gateway for breaches, can we capture the source and circumvent jeopardizing the whole network?”. Lastly, “can we know the topology of the identity network? How is it interconnected or how is every PII located in the Ecosystem?”.

We then map the questions into graph based model as ”In a graph-based network of identity, what are the isolated PII nodes and connected ones? If formed as an interconnected cluster, what are the ones on boundary and on center?”. We would also want to answer ”Inside the network, which PII is in the critical path of obtaining others most often and which PII can influence the acquisition of the other PII?”.

In this thesis, we focus on three statistical indices on the given data set: (1) Bar, pie, and distribution charts based on the node or edge value (2) Centrality measurement including node specific in and out-degree centrality, betweenness centrality, and closeness centrality and (3) Strongly connected components of nodes for identifying clusters. Based on the results, a comprehensive discussion is presented about possible breaches with more important

attributes, and flow of personal information inside the network modeling the real-world information movement.

Chapter 3

Statistical Charts

We present sets of mathematical formula and statistical chart visualization in this chapter. The data source we used is from ITAP in Ecosystem, which contains 627 PII attributes in total, and divides the analyses based on edge or node specific properties. 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 tuples $e_{ij} = \langle i, j \rangle$ 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. Each node A_j is labeled with a Boolean random variable, denoted $D(A_j)$, which is true if the attribute has been exposed and false otherwise. For simplicity, we consider all edges to be independent. Therefore, we can assign conditional probabilities to each edge with $Prob(e_{ij}) = Prob(D(A_j)|D(A_i))$.

3.0.1 Statistical charts based on edges

We implement the pie chart to observe the percentage for PII with/without outgoing edges or with/without incoming edges as shown in Figure 3.1. We can observe that 211 PII attributes (33%) are with incoming edges, while 45 (7%) are with outgoing edges.

Insight: Only 7% of PII have any effect on the risk of exposure of others and a total of 33% could possibly be affected. The PII with outgoing edges should be carefully protected. The most important PII in this list are discussed shortly.

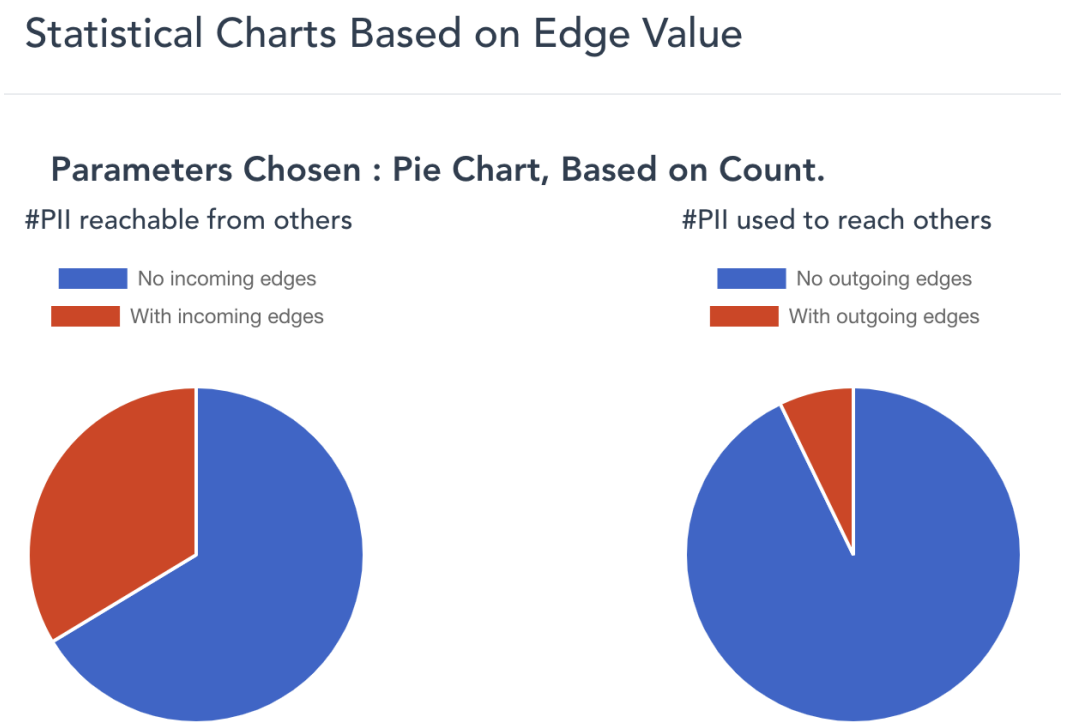

Figure 3.1: A snapshot showing pie charts for percentage of nodes with/without in/out degree.

Furthermore, taking the probability on the edges into account, we extend the degree centrality from summing discrete edge count to accumulating risk. Degree centrality equals the number of links that a vertex has with other vertices. The equation for this measure is as follows:

$$C_{D_{out}}(v_i) = outdegree(v_i) = |\{e_{ij}\}| \tag{3.1}$$

$$C_{D_{in}}(v_i) = indegree(v_i) = |\{e_{ji}\}| \tag{3.2}$$

If we consider the weight (i.e., probability) on edges, this yields the equation:

$$C_{W_{out}}(v_i) = \Sigma Prob(e_{ij}) \tag{3.3}$$

$$C_{W_{in}}(v_i) = \Sigma Prob(e_{ji}) \tag{3.4}$$

Figure 3.2 presents the top 10 PII in descending order based on the number of incoming and outgoing edges. The top three attributes with the highest number of incoming edges, i.e., most easily discoverable through incoming edges, are *Name*, *Credit Card Information*, and *Date of Birth*. Also, the top three attributes with the highest number of outgoing edges, i.e., most likely able to reach the wide variety of PII through outgoing edges, are *Customer Database*, *Password*, and *Email address*. Figure 3.3 shows the same statistics on the top 10 PII with most incoming and outgoing edges, with the difference that it considers the sum of weights on the edges instead of merely the edge count.

Insight: *Name* has the highest rank among PII discoverable from others through incoming edges and *Customer database* sits at the top of nodes with the highest outgoing degree, whether the edge count or edge weight is considered.

3.0.2 Statistical charts based on nodes

- Distribution Chart based on node risk and value

We examine the distribution based on risk and value of each attribute to better understand the underlying trend for all properties. The chart is calculated by fixing linear interval size on x-axis and counting the number of PII lying in each interval. Figure 3.4 gives a snapshot of the distribution chart for node value with interval unit of 100,000 in US Dollar value. According to the ITAP project[27], ITAP determines the loss value of a PII by averaging out the identity theft cases in which the PII was breached as a source of entry. Since ITAP usually lacks the number of victims involved in a case, the loss value is not per victim. Figure 3.5 yields a result for node risk with interval size 0.001.

Insight: The vast majority of PII are valued at less than $100,000 but have a risk of exposure of less than 0.001 too.

- Scatter plot of Closeness vs. Betweenness Centrality

Freeman [10] developed a set of measures for centrality based on betweenness. Later on, he proposed four core criteria, which developed into degree, closeness, betweenness, and eigenvector centrality [9]. We further leverage the concept of closeness and betweenness centrality to investigate the Ecosystem graph.

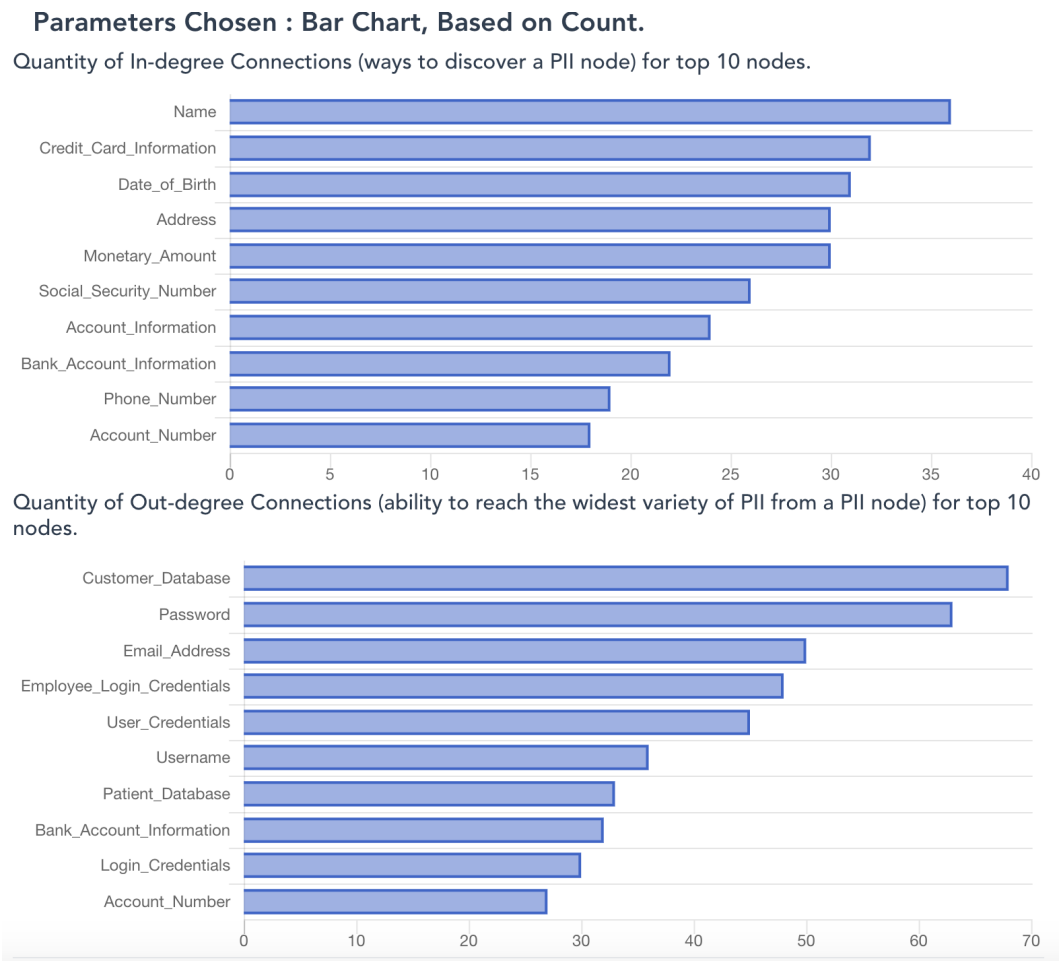

Figure 3.2: A snapshot showing top 10 PII with most in and out degree count.

Closeness Centrality emphasizes how close a vertex is to all other vertices in the topology – the distance of a vertex to all others in the network by focusing on the geodesic measurement from each vertex to all others [9]. To be more specific, it calculates the shortest path between all nodes and assigns each node a score based on the length of its shortest paths to other nodes. According to Yin et al. [5], closeness is an evaluation for "how long it will take

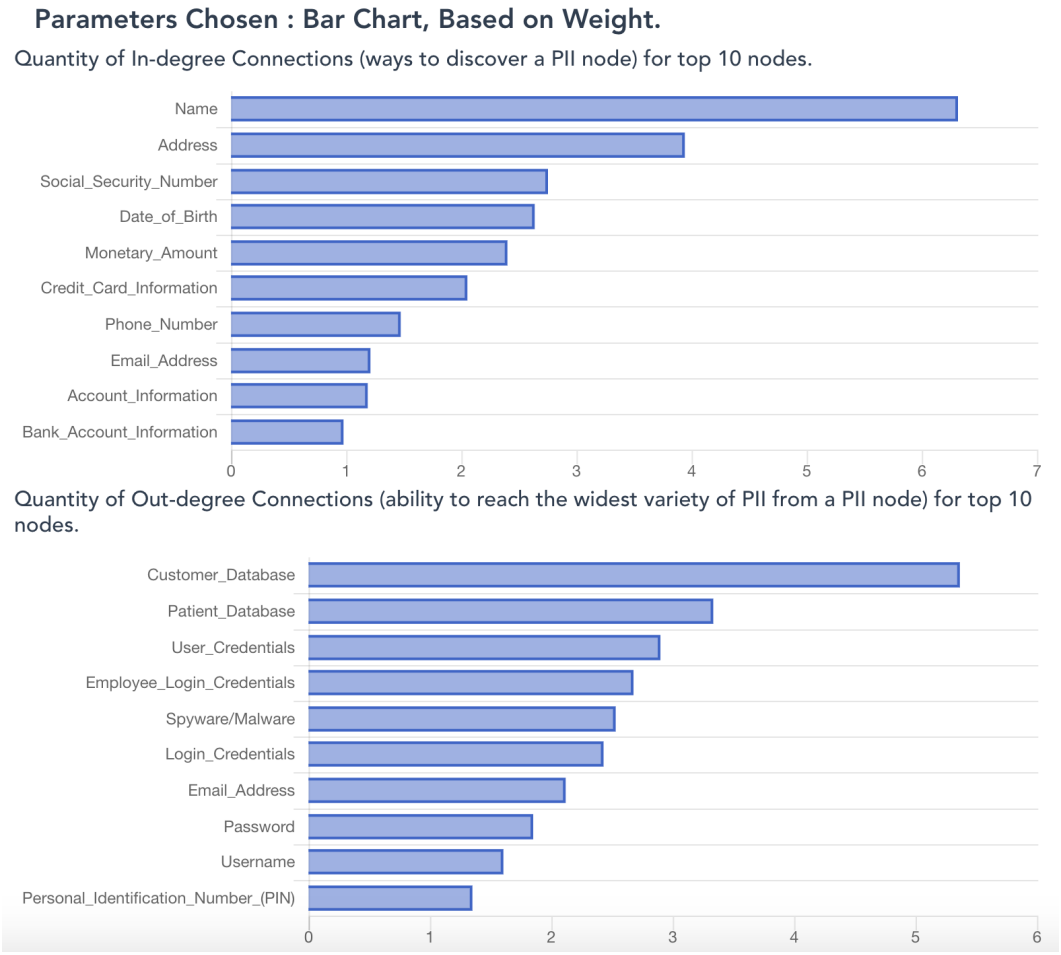

Figure 3.3: A snapshot showing top 10 PII with most in and out probability sum on edges.

information to spread from a given vertex to others in the network" (p.1603), which helps find the PII attributes that are best placed to reach others once breached, and thus influence the entire network most efficiently. Consequently, closeness centrality in the identity ecosystem is a measure of **Information Acquisition Power**. The higher it is for a PII attribute, the more power

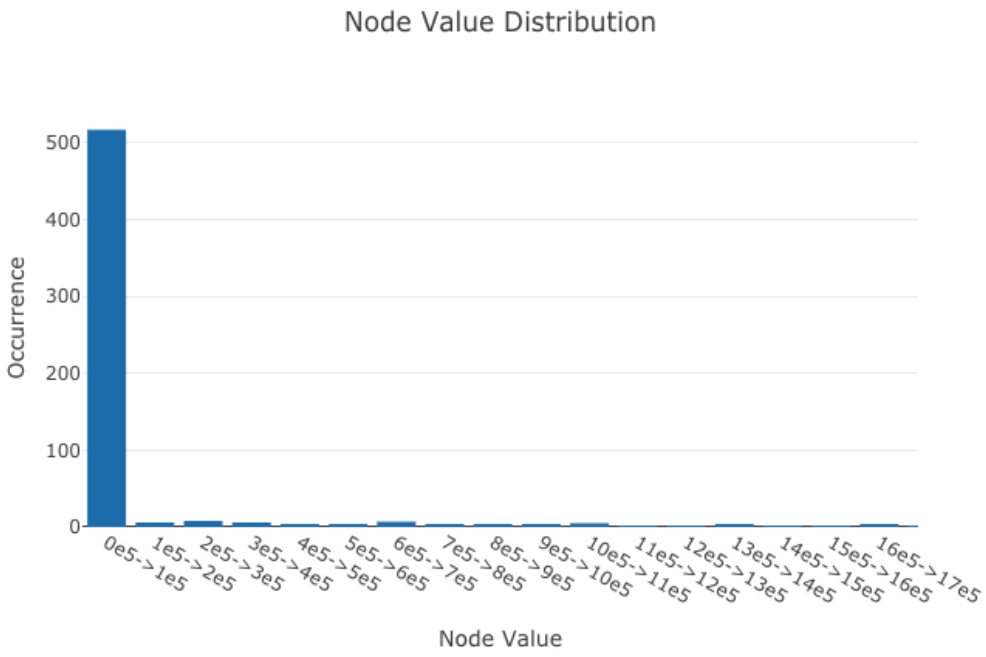

Figure 3.4: Distribution chart based on node value with interval size of $100,000.

that PII attribute has in exploiting the entire identity ecosystem. Such PII attribute would only need few others to discover the whole network. Also according to Freidkin [11], closeness centrality represents the independence in the sense that PII attributes with higher closeness centrality do not need to seek information from other more peripheral PII attributes. This yields the equation as follows. $C_c(v_i)$ stands for the closeness centrality for vertex i and $D(i,j)$ is the distance of the shortest paths between two vertices v_i and v_j (considering the number of edges and not edge weight):

$$C_c(v_i) = \sum_{j=1}^{n} \frac{1}{D(i,j)} \tag{3.5}$$

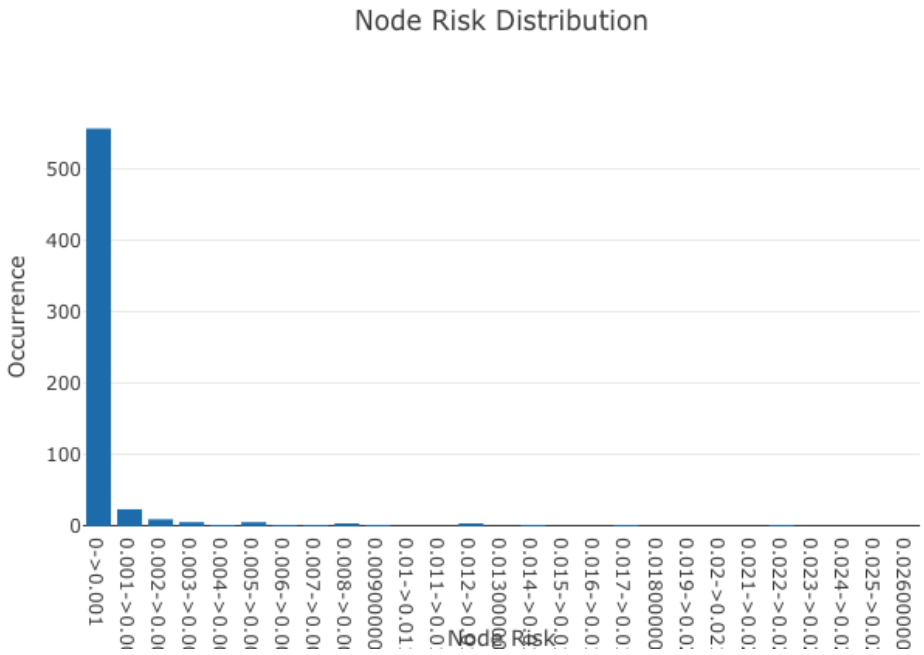

Figure 3.5: Distribution chart based on node risk with interval size of 0.001.

Betweenness Centrality Betweenness centrality [9] serves as an alternative concept of centrality focusing on control over the connections between other pairs of vertices. Betweenness centrality does this by identifying all the shortest paths and then aggregating how many times the node lies on one. Using $\alpha(i, j)$ as the number of different shortest $\langle i, j \rangle$ paths, and $\alpha(i, u, j)$ as how many times the shortest path flows through u ($u \neq i, j$), the equation is as follows:

$$C_B(u) = \sum_{i \neq j \neq u} \frac{\alpha(i, u, j)}{\alpha(i, j)} \tag{3.6}$$

Betweenness centrality recognizes nodes that act as 'bridges' among whole and assesses the PII attributes that determine the flow around the sys-

tem. Betweenness serves as a powerful characteristic for communication dynamics – a high betweenness index could imply a node regulates collaboration in-between, holds authority over, or infers periphery of diverse clusters. In our Ecosystem context, it measures how often a PII attribute is in the critical path of acquiring or discovering other PII, and hence measures **Criticality**.

We calculate the scatter plot of Information Acquisition power (y-axis) vs. Criticality (x-axis). This plot could further be divided into four quadrants based on the combination of high and low values on x and y axes. Denoting C for Criticality (i.e., betweenness) and I for Information Acquisition Power (i.e., closeness), Figure 3.6 shows the graph with blue dots representing high C and high I, Figure 3.7 with green dots low C and High I, Figure 3.8 with red dots High C and low I, and lastly Figure 3.9 with orange dots low C and low I[1].

Insight: Most of the data points maintain low information acquisition power and low criticality (Figure 3.9). There exists only few sparsely distributed data points, discussed in more details shortly, with both high criticality and high information acquisition power (Figure 3.6). Such PII attributes are powerful in acquiring other PII and act as critical bottlenecks in the network of PII attributes. If evaluated using the Ecosystem model, these PII attributes could be asserted as attributes that will rapidly jeopardize the remaining sub-network if already exposed, and boost the information flow of

[1]Low and high are indicating below and above average, respectively.

exposure inside the system. Interestingly, there is only one data point with high criticality but low information acquisition power (Figure 3.8) and that is *Signature*.

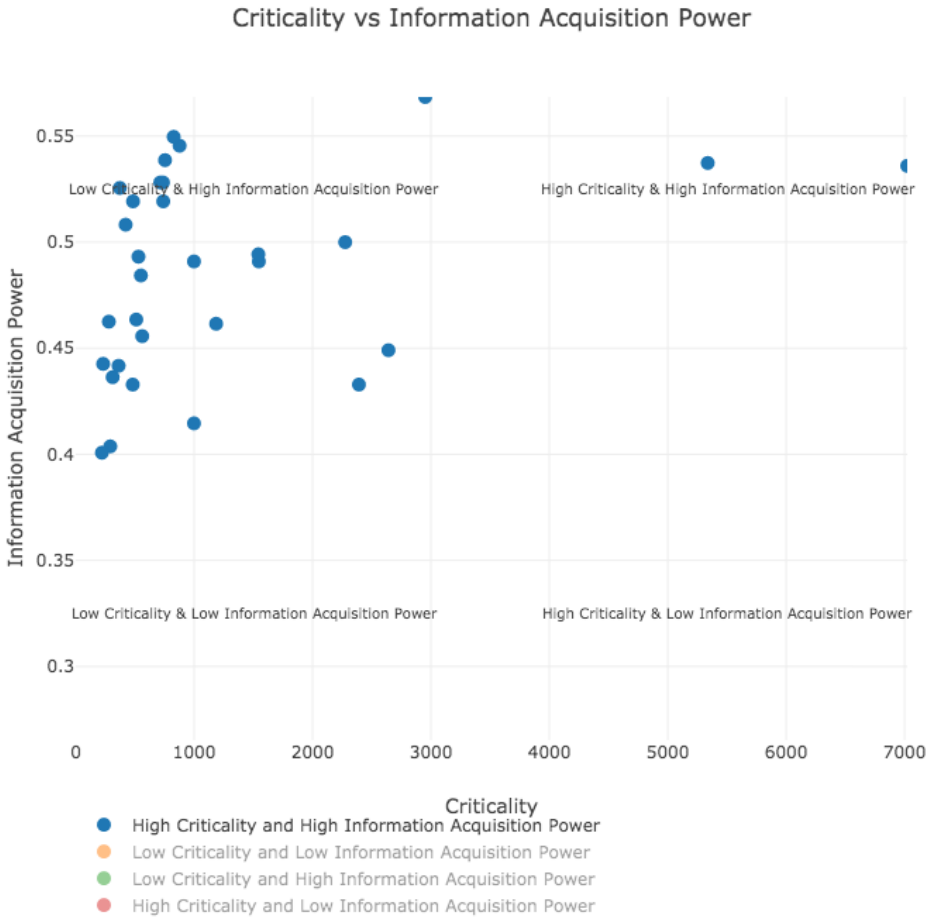

Figure 3.6: Scatter plot with high betweenness (criticality) and high closeness (information acquisition power).

Figure 3.10 displays the top 10 PII in descending order based on the value of information acquisition power and criticality.

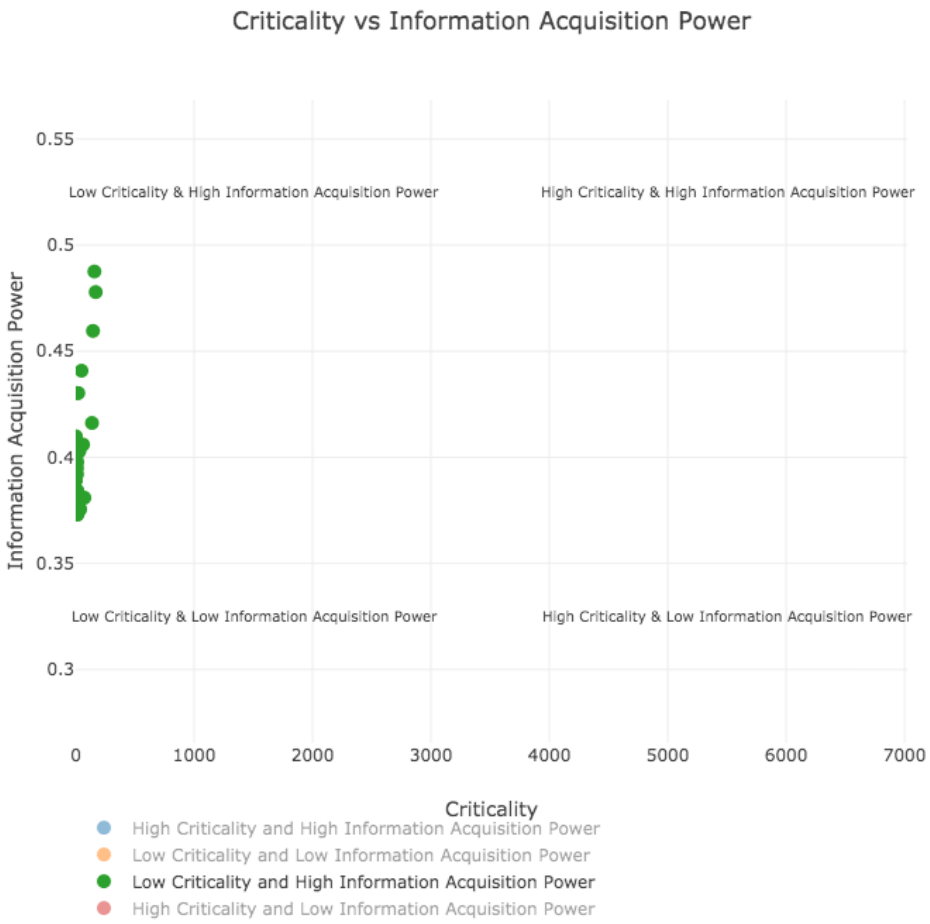

Figure 3.7: Scatter plot with low betweenness (criticality) and high closeness (information acquisition power).

Insight: The top three attributes with the highest value of information acquisition power are *Email Address*, *Name*, and *Address*. The top three attributes with the highest value of criticality are *Customer Database*, *Password*, and *Email Address*.

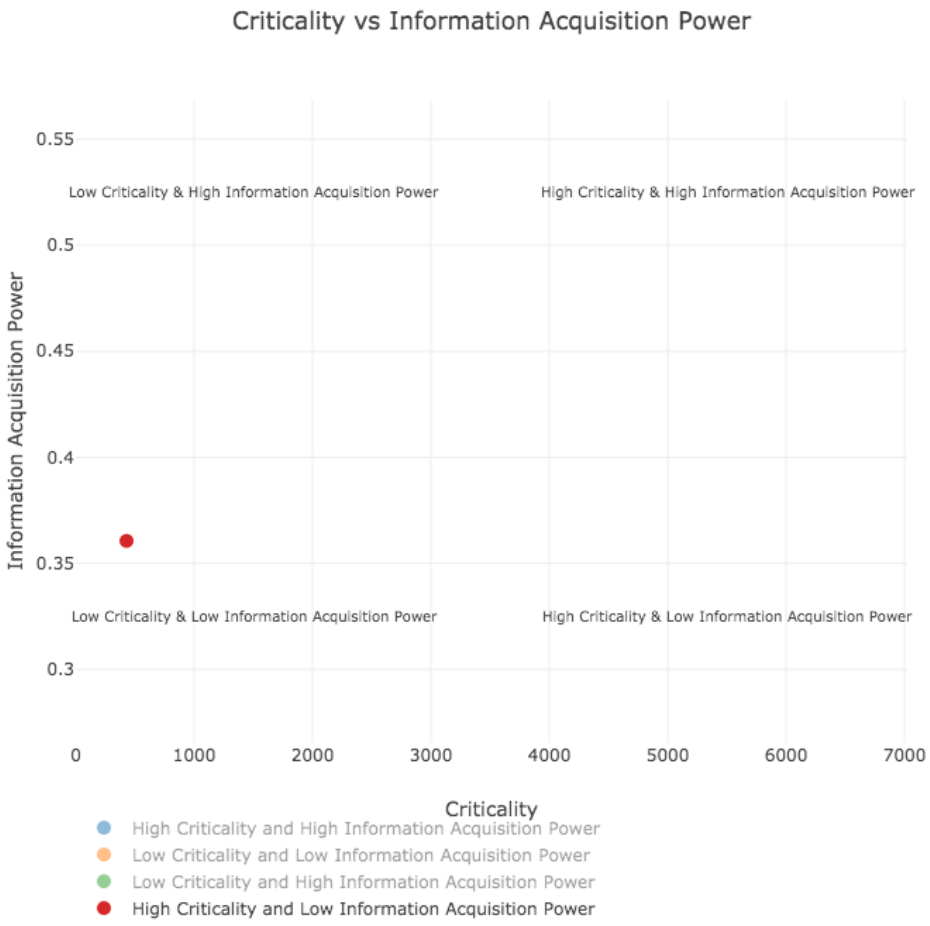

Figure 3.8: Scatter plot with high betweenness (criticality) and low closeness (information acquisition power).

3.0.3 Strongly Connected Components

In the current Identity Ecosystem, a large portion (%65) of nodes is completely isolated from the rest of the Ecosystem. Among those PII attributes with connections, we further identify attributes that are mutually

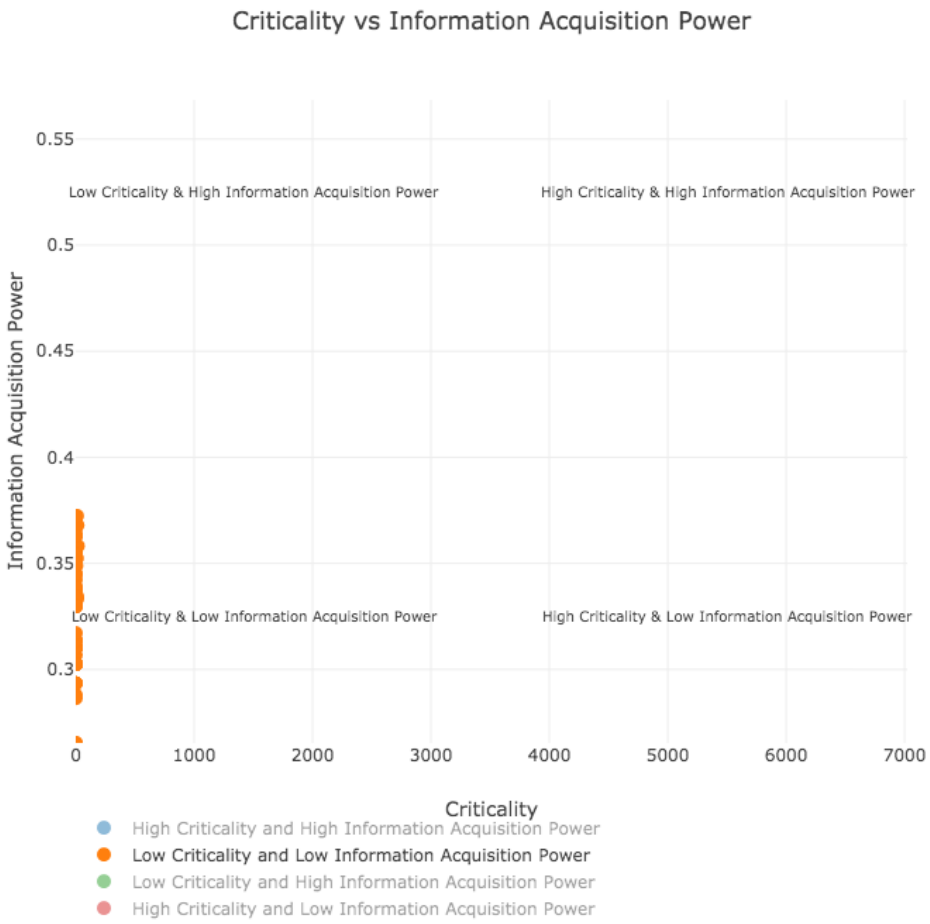

Figure 3.9: Scatter plot with low betweenness (criticality) and low closeness (information acquisition power).

coupled among themselves, which we define as 'clusters'. Clusters serve as subsets that are dangerous sources for breaches, can quickly jeopardize other members in the group and confine the flow inside sub-network.

We propose the cluster to be a Strongly Connected Component (SCC)

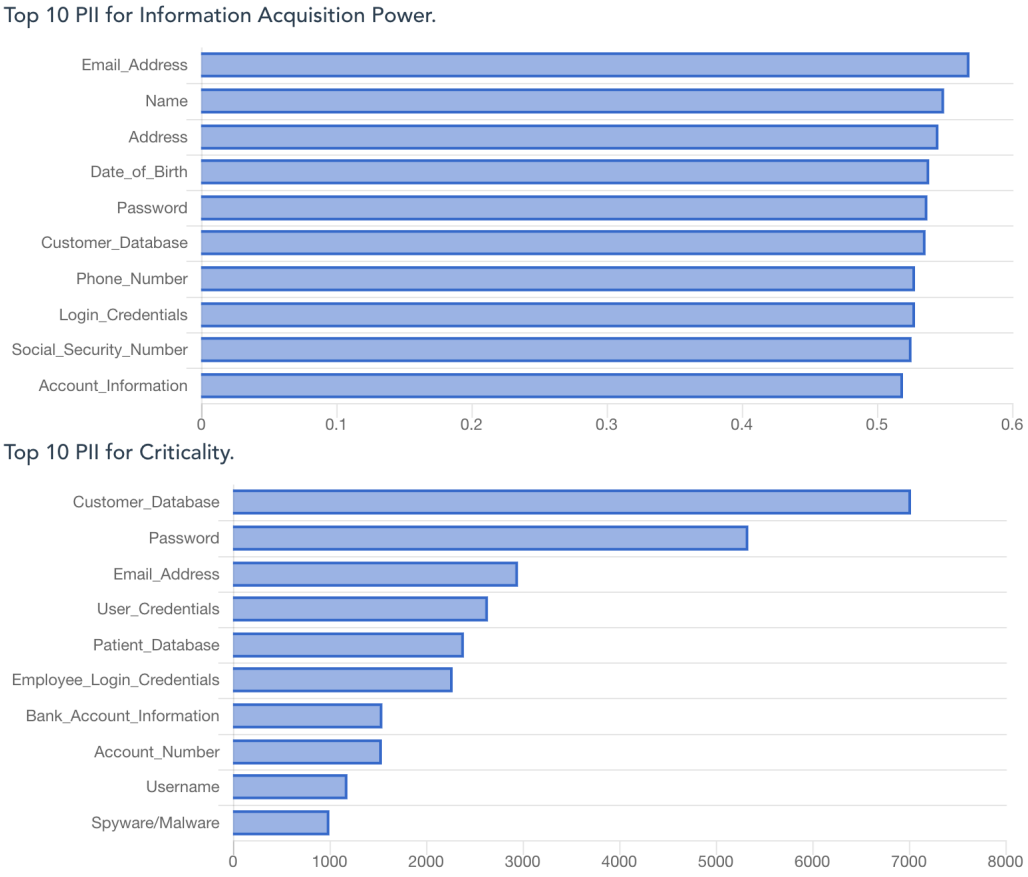

Figure 3.10: A snapshot showing top 10 PII with highest information acquisition power and criticality values.

in the graph theory. A SCC of a directed graph $G = (V, E)$ is a maximal set of vertices $U \subseteq V$ such that for every pair of vertices u and v in U, both $u \mapsto v$ and $v \mapsto u$ hold, where $u \mapsto v$ means there is a directed path from u to v. Consequently, in a cluster, there is a probability that every PII attribute can reveal every other PII and be revealed by every other PII. Tarjan's classic serial algorithm for detection of SCCs runs linearly with respect to the number of

Cluster of attributes, containing 36 vertices sorted in alphabetical order.	
1. Address	2. AccountNumber
3. AccountInformation	4. Age
5. BankAccountInformation	6. BankAccountNumber
7. BiographicData	8. BirthCertificateInformation
9. CreditCardInformation	10. CreditCardNumber
11. CVVCode	12. CheckInformation
13. DateofBirth	14. DebitCardInformation
15. Driver'sLicenseNumber	16. Driver'sLicenseInformation
17. Date	18. EmployeeLoginCredentials
19. EmailAddress	20. EmployeeRecord
21. ExpirationDat	22. IDCardInformation
23. LoginCredentials	24. Name
25. Password	26. PersonalIdentifiableInformation
27. PhoneNumber	28. PersonalIdentificationNumber
29. PhysicalAddress	30. PassportInformation
31. Photograph-Person	32. PatientMedicalRecord
33. RoutingNumber	34. SocialSecurityNumber
35. Username	36. W-2FormInformation

Table 3.1: List of attributes in SCC.

edges and uses depth-first search. We apply Tarjan's algorithm [19] to compute the clusters. We found one cluster of 36 nodes which we display in Table 3.1.

Insight: Every PII in Table 3.1 has a probability of exposing every other PII in that table.

Chapter 4

Discussion of Results

In this chapter, we analyze the statistical results and give example take-aways from the above charts. Considering ITAP model, it filters out edges with very low probability aiming to eliminate noise. Overall the Ecosystem contains 627 vertices. We can observe that a large portion of the nodes is not connected to any other node. In fact, 65% of the nodes are fully isolated without any inbound or outbound connections. Only a small portion is considered to be important when breached or compromised, and one should make an effort to protect them. Among those with connections, we further observe the ranking by degree centrality to speculate candidates with most in-degree versus most out-degree, which could be interpreted as attributes that are most likely to get compromised, versus attributes that tend to spread information.

We further discover possible layout and structural features for the Identity Ecosystem graph by computing the SCC of the network. We extracted clusters, where each node is inter-reachable inside the sub-graph. Between that 33% with incoming and 7% with outgoing edge PII nodes, an overlap of 36 (about 5%) vertices constitutes a big component.

We can assert our ecosystem model to be a sparse graph, where most

attributes are unreachable. Only 5% congregate together and serves as a central concern for our identity management.

We utilized closeness and betweenness centrality to better understand the influence in the topology. Closeness, or information acquisition power in this context, measures the ability of a PII attribute to retrieve information from and send information to others. Those PII attributes with high value can be viewed as 'broadcaster' or 'gossiper', which if breached, can put others in danger. Betweenness, or criticality in this context, is based on the assumption that a PII attribute may be exposing others if it presides over a path bottleneck. It also identifies the boundary spanner, which separates different communities and features. Those PII attributes with high value can be viewed as 'bridge' or 'broker', if one connecting component is breached, those can function as essential endpoints to protect the identity by not allowing information to flow through.

Generally, previous studies indicate that centrality metrics are positively correlated [21] [15]. Overall degree and closeness were strongly intercorrelated, while betweenness remained relatively uncorrelated with the other measures [2]. Combinations of centrality values represent certain topology and positional patterns ([7] p. 51). Given attributes with high degree and low closeness centrality (information acquisition power), we can assert that the PII is embedded in the cluster and far away from others, whereas low betweenness (criticality) infers that the PII holds redundant links where information just bypass it. Given attributes with a low degree and high closeness centrality

(information acquisition power), the PII ties with substantial or active others, whereas high betweenness (criticality) indicates that PII is spanning few links, but with crucial influence on network flow. Low closeness (information acquisition power) and high betweenness (criticality) combination results in specific PII monopolizing the ties from a small number of PII attributes to many others. We found a prime example of such situation with *Signature*. Low betweenness (criticality) and high closeness (information acquisition power) portray the PII locates in a dense, active cluster at the center of events with many others. We summarize different combinations and their corresponding characteristics in Table 4.1.

	Low Degree	Low Closeness (Information Acquisition Power)	Low Betweenness (Criticality)
High Degree	-	Embed in a cluster which is faraway from others	PII with redundant connection - flow bypass
High Closeness (Information Acquisition Power)	Key PII connected to important and active others	-	Center PII located in a dense, active cluster at the center of events with many others
High Betweenness (Criticality)	PII's few ties are crucial to network flow	PII monopolizes the ties from a small number of PII to many others	-

Table 4.1: Combinations of centrality metrics.

Chapter 5

Related Work

In this chapter, we cover previous research that studies and surveys the statistics of identity theft. We can categorize previous work into three main sources: Federal and State agencies, private organizations and academic institutions.

From government sources, Federal and State agencies, studies by U.S. department of Justice (Harrell [14]) release reports on distribution of identity theft victims. Also United States General Accounting Office (USGAO [20]), Federal Trade Commission (FTC [12]), Office of the Inspector General, Federal Bureau of Investigation (FBI), Postal Inspectors Office, and United States Secret Service (USSS) present studies on identity theft from different domains.

Among private organizations, Javelin [16] publishes comprehensive analysis and case studies about fraud detection and identity threat.

In the academia, Copes et al. [6] analyzed reports from National Public Survey on White Collar Crime and summarized financial-related fraudster behavior such as credit card fraud and bank account fraud. Allison et al. [1] gathered data from agencies. They performed statistics analysis on victims and extracted demographic patterns of victims among the general U.S. popu-

lation. Using Routine Activity Theory, Reyns [18] reported an empirical study of identity theft in the United Kingdom. Pratt et al.[17], and Choo, [4] also conducted studies utilizing Routine Activity Theory in different jurisdictions.

In these studies, statistic were presented. However, those data sets were not fully constructed into a structured mathematical model and do not interact with graph theoretic and social network analysis measures. We feed data sets from ITAP and model the risk of exposure using Bayesian Network [25]. We are also one of the first to develop identity ecosystem into graph network and exploit the concept from three types of centrality as well as strongly connected components.

Chapter 6

Conclusion

In this thesis, we designed and implemented a visualization framework that assists data providers and collectors to comprehend and analyze the probabilistic graphical model of identity attributes. The visualization tool facilitates understanding of the whole risk model. Based on the Bayesian network presentation of identity attributes, we developed traditional statistical charts such as histograms, scatter plots, and pie charts based on values for each PII to inspect the underlying distribution. Even though hundreds of PII constitute the whole system, a large amount is indeed isolated. Only a small portion of the PII is vulnerable to identity theft and one should make an effort to protect them.

To investigate the structural topology and correlation between PII, we further proposed to apply centrality measures such as degree, closeness, and betweenness centrality. Moreover, we discussed the combination of all the three centrality measures with high and low values. With these measures, we can estimate the hidden characteristics of the network.

Lastly, we calculated Strongly Connected Components (SCC) to recognize clusters of PII that are mutually reachable among themselves. SCCs

are subsets which are dangerous origins for breaches, can quickly jeopardize other PII in the group and constraint the flow inside the sub-network. In the current Identity Ecosystem, there is only one big cluster with 36 PII (5% of the entire ecosystem) interconnected. We can again confirm that as complex as the Identity Ecosystem is, a small portion is considered most threatening and risky.

As the ITAP project continues to collect data, theories and technologies developed from this research can be customized along the way to minimize our identities' risk of exposure and maximize privacy.

Bibliography

[1] Stuart F.H. Allison, Amie M. Schuck, and Kim Michelle Lersch. Exploring the crime of identity theft: Prevalence, clearance rates, and victim/offender characteristics. *Journal of Criminal Justice*, 33(1):19 – 29, 2005.

[2] John M. Bolland. Sorting out centrality: An analysis of the performance of four centrality models in real and simulated networks. *Social Networks*, 10(3):233 – 253, 1988.

[3] S.K. Cherivirala, F. Schaub, M.S. Andersen, S. Wilson, N. Sadeh, and J.R. Reidenberg. Visualization and interactive exploration of data practices in privacy policies. In *Symposium on Usable Privacy and Security*, pages 3–10, 2016.

[4] Kim-Kwang Raymond Choo. The cyber threat landscape: Challenges and future research directions. *Computers and Security*, 30(8):719 – 731, 2011.

[5] Li chun Yin, Hildrun Kretschmer, Robert A. Hanneman, and Ze yuan Liu. Connection and stratification in research collaboration: An analysis of the collnet network. *Information Processing and Management*, 42(6):1599 – 1613, 2006. Special Issue on Informetrics.

[6] Heith Copes and Lynne M. Vieraitis. Understanding identity theft: Offenders' accounts of their lives and crimes. *Criminal Justice Review*, 34(3):329–349, 2009.

[7] Donglei Du. Social network analysis: Centrality measures, March 2019.

[8] Lorrie Faith Cranor. Platform for privacy preferences (p3p). *Encyclopedia of Cryptography and Security*, pages 940–941, 2011.

[9] Linton C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 40(1):35–41, 1977.

[10] Linton C. Freeman. Centrality in social networks conceptual clarification. *Social Networks*, page 215, 1978.

[11] Noah E. Friedkin. Theoretical foundations for centrality measures. *American Journal of Sociology*, 96(6):1478–1504, 1991.

[12] Federal Trade Commission (FTC). 2016. consumer sentinel network data book for january to december 2016, March 2016.

[13] Gemalto. Mining for database gold: Findings from the 2016 breach level index, March 2016.

[14] Erika Harrell. Victims of identity thef, 2016, 2016. `https://www.bjs.gov/content/pub/pdf/vit16.pdf`.

[15] Natarajan Meghanathan. Correlation coefficient analysis of centrality metrics for complex network graphs. In *Computer Science On-line Conference*, pages 11–20, 01 2015.

[16] Kyle Marchini Pascual, Al and Sarah Miller. 2018 identity fraud: Fraud enters a new era of complexity, March 2018.

[17] Travis C. Pratt, Kristy Holtfreter, and Michael D. Reisig. Routine online activity and internet fraud targeting: Extending the generality of routine activity theory. *Journal of Research in Crime and Delinquency*, 47(3):267–296, 2010.

[18] Bradford W. Reyns. Online routines and identity theft victimization: Further expanding routine activity theory beyond direct-contact offenses. *Journal of Research in Crime and Delinquency*, 50(2):216–238, 2013.

[19] R. Tarjan. Depth-first search and linear graph algorithms. In *12th Annual Symposium on Switching and Automata Theory (swat 1971)*, pages 114–121, Oct 1971.

[20] USGAO. 2017.services offer some benefits but are limited in preventing fraud, March 2017.

[21] Thomas Valente, Kathryn Coronges, Cynthia Lakon, and Elizabeth Costenbader. How correlated are network centrality measures? *Connections (Toronto, Ont.)*, 28:16–26, 01 2008.

[22] Kami Vaniea, Qun Ni, Lorrie Cranor, and Elisa Bertino. Access control policy analysis and visualization tools for security professionals. *SOUPS Workshop (USM)*, pages 7–15, 04 2019.

[23] Verizon. 2018 data breach investigations report, March 2018.

[24] Robert W. Reeder, Patrick Kelley, Aleecia M. McDonald, and Lorrie Cranor. A user study of the expandable grid applied to p3p privacy policy visualization. In *Proceedings of the 7th ACM workshop on Privacy in the electronic society*, pages 45–54, 01 2009.

[25] 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 *2016 IEEE International Carnahan Conference on Security Technology (ICCST)*, pages 1–8, Oct 2016.

[26] Razieh Nokhbeh Zaeem, Monisha Manoharan, Yongpeng Yang, and K. Suzanne Barber. Modeling and analysis of identity threat behaviors through text mining of identity theft stories, 2017.

[27] Jim Zaiss, Razieh Nokhbeh Zaeem, and K. Suzanne Barber. Identity threat assessment and prediction. *Journal of Consumer Affairs*, 53(1):58–70, 2019.

Vita

Chia-Ju Chen received the Bachelor of Science degree in Electrical Engineering and Computer Science from National Tsing Hua University, Taiwan and continued to pursue Master of Science degree in Electrical and Computer Engineering in University of Texas at Austin.

Permanent address: ju40268@utexas.edu

This thesis was typeset with LaTeX[†] by the author.

[†]LaTeX is a document preparation system developed by Leslie Lamport as a special version of Donald Knuth's TeX Program.