Data is growing exponentially, .5% of all data is analyzed and the amount is decreasing. [trend] To address this technology is changing and more investments in modern data infrastructure are being made. The investments are to improve data analytics, which includes real-time data processing and visualization. As data increases and technology advances, the types and severity of attacks will vary. Privacy and sensitivity of the attributes will contribute to the severity of a data breach. This research analyzes the impact of different privacy metrics on the severity of a data leakage. A model of weighing different data classifications is proposed.

Severity is measured by the number of records have been leaked. The ability to identify an individual is taken into consideration of a breach’s severity. The impact of a data leakage on an individual’s life should outweigh the number of records that was leaked when measuring severity. For example, 10 records that reveal patients have the cold may be considered less severe than 5 records that reveal patients with HIV. The existing conditions when the leak occurred may also impact severity. Records leaked maliciously can remain undetected longer, which increases the impact on an organization. Other factors that increase severity are linkages and frequency of being the victim to an attack. Linkages are the relationships that are publically available that can be used to reveal identities.

**Regulations** have been increased to protect the confidentiality of users. The European Union Agency for Network and Information Security (ENISA) handles information and network security throughout the European Union (EU). Guidelines set by ENISA are to Prepare, Detect, Notify and Respond to a security incident. [enisa] Accuracy and severity must be measured when handling confidential user information. Data should remain anonymous and be generalized. Consent may be mandatory and provided with an “Opt-out” choice. Data handling must be done with intent to fulfill a purpose.

Many variables are involved when measuring the severity of a data breach. The criticality of the data can be determined on sector-based analysis. For example, a breach of confidential data of an organization can negatively impact their stock price. [business] [Biz] et al. saw a 5% decrease in stock price when a company is a victim of a confidential data breach. Data breaches not involving confidential information had no effect. Organizations lost 2.1% of their market value within 2 days of disclosure. [CHECK business]

Sweeney et al. proposed k-anonymity, which enforces quasi-identifier values occur k number of times. K-anonymity makes uniquely identifying individuals more difficult. Often generalized data is released and unintentionally discloses information about individuals. These attacks are done with numerous techniques such as inferring knowledge or linking attributes to other data sources. 87% of individuals can be identified by their 5-digit zip code, gender and date of birth. [Sweeney] The government and medical industries commonly release information containing these attributes unaware of different linkage and inference attacks. Security of data can protect against a direct data breach, but not from information leaked through inference. Data can be leaked through multiple queries. Queries individually may pose little risk, but when used together, an attack can infer sensitive information.

These attacks and vulnerabilities are not new within the industry or research. Statistical databases are released to provide data for research in data mining or fraud detection. A technique to generalize the data involves adding noise, which can damage the integrity of the information. Multi-level databases (MDB) stores data in different classifications. For example, data is split up into higher and lower classified information. A vulnerability that can be found is when lower classified information can be used to infer higher classified records. A way to mitigate this vulnerability is through strong database design. However, the replication of data after it is released poses a risk. After data leaves the original source, it can be copied and manipulated many times after. There is little to no control over the handling of the data once the data has reached multiple receivers. To avoid this vulnerability, all sensitive data can be suppressed, but this technique can leave the data useless.

Sweeney et al. proposed k-anonymity to address some of these issues. However, k-anonymity is not perfect and is susceptible to different attacks. An unsorted matching attack occurs when positions of the tuples in each generalized table match the private table. To prevent this attack, randomly sorting the data is necessary. Complementary released table vulnerability is when two generalized tables form a linked table. This linked table then uses the combined quasi-identifier values to uniquely identify rows in the private table. The vulnerability can be addressed by using the quasi-identifiers of the original table. Another technique is to base the new generalized table after the original table that was released. When basing the new generalized table from the original table, no value should be more specific than the original table. For example, if the original table generalized their zip codes to 0213\*, the new table should not be more specific with 02139. The final attack the Sweeney et al. identifiers is the temporal attack. Temporal attacks occur when new data is added to the private table over time and a new generalized table is released. Linking the original released table with the newly released table can reveal unique rows. A way to avoid this is to base the newly released table on the original released table plus the newly added information.

Ashwin Machanavajjhala et al. presents two attacks on *k*-anonymity, homogeneity attack and background information. An attacker can discover sensitive attributes when the data is not diverse enough. A homogeneity attack leaks information due to the lack of diversity in the sensitive attribute. An attacker may have background knowledge, which can be used to infer knowledge. Ashwin Machanavajjhala et al. proposes Bayes-Optimal and *l*-diversity. Bayes optimal is an algorithm that works under the assumption that the data publisher and adversary know the complete distribution set of sensitive and non-sensitive attributes. *L*-diversity provides privacy without the data publisher knowing how much background information an adversary may have. Although Bayes optimal covers a wider scope, it is not practical in use. It is unlikely that the adversary and data publisher have complete knowledge of all sets of sensitive and non-sensitive attributes.

Each block of quasi-identifier groups, q-blocks, should have the same frequency of sensitive attributes. The frequency of sensitive attributes can protect against knowledge an attacker may know. Ashwin Machanavajjhala et al. proposes 2 versions of *l­­*-diversity, entropy and recursive diversity. Entropy diversity ensures that q-block has well represented groups of sensitive attributes. The more uniform a q-block is, the higher the entropy. Recursive diversity is another algorithm that checks if a sensitive attribute occurs too frequently. Ashwin Machanavajjhala et al. proposes other algorithms to handle non-sensitive attributes, which involve variations of entropy and recursive diversity.

1. Attacker can discover values of sensitive attributes when data is not diverse enough.
2. Attackers have background knowledge – k-anonymity does nto protect against this
3. Presents two attacks on k-anonymity – Homogeneity attack and background knowledge attack
4. Homogeneity attack can leak information due to lack of diversity in the sensitive attribute. This is the quasi-identifier problem, but in reverse.
5. An example of a background knowledge attack is to use demographics to infer knowledge.
6. Severity vs Privacy\*\*\*\*\*\*\*\*\*\*\*\* Secure vs Privacy – similar problem
   1. Severity is a function of privacy
   2. Privacy is not a function of severity
   3. Measuring severity can provide an idea of the impact with privacy based on that function
   4. Can measure severity better – high profile record involved
7. Contribution Bayes-optimal and L-Diversity
8. Bayes-Optimal is for the assumption that the data publisher and attacker know the complete set of sensitive and non-sensitive attributes
9. L severity provides privacy even though the data publisher is not aware of the background knowledge an attacker may have. L-diversity is the requirement that sensitive attributes are well represented in each group.
   1. L-diversity is considered more practical than bayes optimal
10. Limitations (Why we don’t need it)
    1. Insufficient Knowledge – does not know entire distribution of data
    2. Adversary most likely does not know the distribution either
    3. Instance level knowledge
    4. Multiple adversaries with different levels of knowledge
11. For every generalized block the same frequency of sensitive attributes should be throughout
12. For each q-block there should be L represented sensitive values. This can serve as a measurement to how much background knowledge an attacker must know.
13. Complex – How many unique of S is in each q-block
14. 2 instantiations of L diversity
15. Entropy Diversity – Captures well represented groups. The more uniform a group is the higher the entropy.
16. Recursive Diversity –
17. Must consider non-senstiive attributes, should not separate them out (MEASURE SEVERITY)
18. Positive Disclosure-Recusrive: if we can remove a sensitive attribute without s being too frequent
19. Recursive diversity can be done at a row level. None of the values shuld appear too frequently
20. Entropy vs recursive diversity

Common problem \*\*\*PERFORMANCE

Not enough information?

l-diverse not good enough when presented with background information

L-Severity: DF considered K-anonymity, + L-Diversity

Assumption sensitive attribute

Combination of k-anonymity/l-diversity and data modeling is most ideal situation.

Future research – Focus on don’t care sets, continue with l-diversity

Include t-closeness in research

Experiment phase

0 if n(q, s’) >= l

l-n(q, s’) else

l must be defined

well represented – less severe

Higher the # of sensitive attributes less severe the leak is

CS/DF/# of distinct sensitive attributes

Assumption is that the higher the number of distinct attributes in each block, the less likely it is to identify an individual

Too Frequent

For each s in S

N(q,s)

DivFactor

CS/DF/Div Factor

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Job | City | Sex | Age | Disease | Medication |
| Student | USA | Human | \* | HIV | Vitamins |
| Student | USA | Human | \* | HIV | Vitamins |
| Student | USA | Human | \* | Heart Attack | Paracetamol |
| Student | USA | Human | \* | HIV | Vitamins |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Job | City | Sex | Age | Disease | Medication |
| Student | LA | Male | 10 | HIV | Vitamins |
| Student | LA | Male | 20 | Heart Attack | Aspirin |
| Student | LA | Male | 30 | Cancer | Paracetamol |
| Student | LA | Male | 40 | Meningitis | Aspirin |