Quasi-identifier values can impact the severity of a data leakage. For example, a leakage involving diseases can have different impacts on individuals at certain stages of their lives. For example, a 96 year old with HIV is included in a data breach, which reveals his or her disease to the public. Disclosure of a disease can have varying consequences. Privacy of a minor is critical and it may be considered that information involving a minor may be more critical. Using L-Severity model only sensitive attributes are considered without any separation from quasi-identifiers or non-sensitive attributes. Non-sensitive attributes may be misclassified and can give more information than intended. Therefore, attaching a constant score for any added attribute can be beneficial when detecting for data leaks. A caveat of scoring severity is that the domain of these attributes and classifications must be maintained. If an attribute is misclassified an attac­­­­­­­­ker can target non-sensitive attributes to prevent detection. Tracking transactions within an application can alert an organization at the time a possible breach has occurred. A breach may go undetected until a victim reports a problem or an attacker advertises the data on the black market. **A DLP** may be tracking when sensitive data is released or based on some measurement of severity. Severity is complex and can involve many dimensions that are shared or specific to an industry. However, there are few a publications that focus strictly on the severity of only the data and the definitions of the impact of values for sensitive, quasi-identifier and non-sensitive attributes. Vavilis et al. created a model to quantify severity by attaching severity scores to values within a sensitive domain. L-Severity does not separate different classifications of the data. We propose a model that is more scalable allowing emphasis on other attribute classifications.

Giving a constant score to non-sensitive attributes can be useful when reading unstructured data. For example, an API can accept JSON objects that may have varying properties. Due to the unstructured format that the data can come in, unexpected attributes may be passed. Depending on how this data is used, extra information may be leaked or accidentally disclosed to an unauthorized user. An example could be a data dump of values that need to be emailed to another group or data that is passed into another system. This can cause system errors, rejections and leakages.

The comparison of the impact of the Dependency Factor (DF) in L-Severity was done against k-anonymity. This research did not find a significant impact on the severity when alternating algorithms. For example, if a table is conforming to k-anonymity the DF can also remain constant when. Having a higher DF metric will reduce the severity of a row. However, a higher DF score does not guarantee that a leaked table conforms to the k-anonymity rule. In order for a record to follow the *k-anonymity* rule, it must be part of a group of records that is at least *k* in size. After our analysis of M-Score’s DF metric and *k-anonymity* it was concluded that they are almost equivalent on the impact of the severity score. *K-anonymity* is a good baseline for measuring privacy within a generalized dataset. We attempted to measure the impact of considering how far off a group of records were from being conforming with *k.* The farther the number of unique quasi-identifiers a group of records is from *k,* the higher the severity will be. This correlates with a lower DF metric. For example, if there is only 1 distinguished record out of *n* records, *n/1* is greater than *n/k.* Assuming that *k* is larger than 1, this will result in a higher severity for a given record. Future research should be done on the impact of using *l-diversity*. [Sweeney] *L-diversity* provides privacy without knowledge of what the attacker may know. For example, the attacker may have strong background knowledge of the data.

**Having** the capability to attach weights to different classifications allows for more detailed analysis. For example, it is possible to weigh privacy higher than other data classifications. Classifications can include the traditional sensitive, quasi-identifier and non-sensitive attributes or attributes that are specific to an industry. Normal traffic of an application or system must be defined. Vavillis et al. use examples where data is being queried from a system. The regular use of the system should be defined so when high volumes or sensitive information is being retrieved the information can be flagged and raise an alert to the proper party. DLP technology is correlated with organizations have teams to prevent data theft. [mcafee] McAfee performed a study where 64% of security professional within firms that experienced a data breach agreed that the breach could have been prevented if their firm used DLP technology. DLP technology is a top tool in detecting insider threats. A tool that detects the severity of data being retrieved can be helpful when investigating security events.

Ashwin Machanavajjhala et al. presented two attacks to perform on *k*-anonymous tables. A homogeneity attack can be mitigated by having l-diversified sensitive values for each q-block. With a well-diversified table, an data publisher can still provide privacy with the threat of multiple adversaries with different levels of background knowledge. We present a DivFactor to include in the L-Severity equation. The DivFactor will provide a better metric for providing an accurate severity score.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Job | City | Sex | Age | Disease | Medication |
| Student | LA | Male | 100 | HIV | Vitamins |
| Student | LA | Male | 20 | Heart Attack | Aspirin |
| Student | LA | Male | 30 | Migraine | Paracetamol |
| Student | LA | Male | 40 | Hypertension | Aspirin |

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

In case 1 and case 2 both leaked tables conform to 1-anonymity. All values are equal except for the age in record 1 for case 2. Using the L-Severity model that Vavillis et al. proposes, the severity of the two tables would be equal. This may be due to a fault in the proposed scores.

|  |
| --- |
| Quasi-Identifier Scores |
| *f(age < 10) =* .8 |

To remain consistent with previous research we attached a constant value for quasi identifier scores of .1. Using the above rule, Case 2 will now be the more sensitive table. Although there is a graphical tree model that depicts the sensitive attributes, it was not included in Vavillis’ calculations or equations. Previous research does not provide support for weighing different data classifications. In Case 1 and 2 all values are the same except for one record’s age attribute. When handling tables with different information, weighing a data classification more than others may return different results.

The L-Severity equation does not consider susceptibility to a homogeneity attack. Assuming Cancer and Meningitus is given the same severity score as HIV, using L-Severity, the severity for each table would be equal. If the number of distinct values within each q-block is considered, table 2 would have less severity.

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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Job | City | Sex | Age | Disease |
| Student | USA | Male | \* | HIV |
| Student | USA | Male | \* | Heart Attack |
| Student | USA | Male | \* | Cancer |
| Student | USA | Male | \* | Meningitis |

We present a Diversity Factor (DivFactor), which counts the number of distinct sensitive values within each q-block. The more diverse the sensitive values are the harder it would be to identify an individual. For example, knowing based on one’s ethnicity that the risk of a heart attack is low. An adversary can assume or infer the individual’s disease. Table 1 has 2-diversity, if Alice assumes Bob is at low risk of a heart attack and can infer that Bob has HIV. Using Table 2, Alice would rule out heart attack, but would still have to figure out whether Bob has HIV, Cancer or Meningitis.

1. No difference between DF and k-anonymity
   1. # of distinct quasi-identifiers
   2. Lower the number the higher the severity
   3. Can replace with k-anonymity, but it will either be generalized or will have same correlation as DF
      1. Generalized: Can perform at the table level or weigh it at the q-block level
      2. If we go by how far off the number of q tuples is from k anonymity, it will have the same correlation as tracking DF
2. Including DivFactor is beneficial
   1. Nothing in place considering the diversity of sensitive values in q-block
   2. L severity is important
      1. Different adversaries with different levels of knowledge
      2. Provides privacy with an attacker having background knowledge
         1. Protects against instance level knowledge
   3. Severity can identify when generalized tables are out of sync with k-anonymity.
   4. A well-diversified table with k-anonymous tables can protect against linkage attacks.
   5. Alerts
3. Weighing different classifications is beneficial for analysis
   1. System perspective
4. Provided an analysis of L-Severity incorporating k-anonymity and l-severity
   1. Analysis of pros and cons and ways of incorporating each technique into measuring the severity of a leaked table
5. Future research
   1. T-closeness
   2. Frequency of sensitive attributes
   3. Consideration of non-sensitive attributes
6. Combination of privacy metrics will provide best results
   1. Q-blocks using k anonymity
   2. Diversity using L diversity
   3. L-severity for impact
7. Severity considers
   1. Privacy
   2. Domain specific knowledge