Data is growing exponentially, .5% of all data is analyzed and this number is shrinking. [trend] To address this technology is changing and more investments in modern data infrastructure are being made. These investments are to improve data analytics such as 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 analyses current measurements of severity and proposes a extended model of L-Severity that includes other dimensions such as context.

Severity is often measured by the number of records have been leaked. The ability to identify an individual is taken into consideration of a breach’s severity. Measuring the severity of a data breach based on the impact it will have on the lives of the individuals and organization can outweigh the impact of the number of rows that were obtained. For example, 10 records that reveal patients have the cold may be considered less severe than 5 records that reveal patients with HIV. The context of the leak 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. Linkages are the relationships that are already publically available that may relate to a leaked or generalized table. Frequency of an attack can also contribute to the severity of a data breach.

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.