Understanding the Effect of Data Quality on Analysis Results

Bhowmick, Sanjukta

The success of artificial intelligence(AI) and machine learning (ML) techniques is dependent on the quantity and quality of data. Despite the promise of data deluge, in reality, it is very difficult to obtain large amounts of representative data. This is due to many factors including, (i) the cost of gathering data, (ii) inconsistencies and bias in reported data and (iii) privacy issues dealing with data sharing. The technological, social and legal constraints of data gathering are unlikely to change soon. Thus, a crucial challenge for AI/ML researchers is to determine the how to deliver accurate results in face of these constraints.

Although this problem is prevalent in all areas of AI/ML, I will specifically discuss in the context of networks analysis. Networks (or graphs) are mathematical models of complex systems of interacting entities. Analysis of the network structure provides insights to the properties of the underlying system. Network analysis is used in many disciplines from identifying genes with similar functions in gene correlation networks to recommendation systems in online markets. We consider issues with respect to these three types of poor-quality data;

Incomplete Data. This is data where parts of the information, here edges or vertices are missing. The challenge lies in either filling in the missing information—such as through link prediction or in evaluating by how much the missing data affects the results. Some of our preliminary results show that the *accuracy is determined by not so much as the number of edges missing, but by the position of edges in the network*.

Erroneous/Biased Data. Erroneous data is distinguished from incomplete data, in that spurious edges/vertices that were not in the original network may have been added. Erroneous data can be detected by gathering the data multiple times, and comparing the associated network models. However, this is an extremely time consuming process, and often not feasible.

An analytical approach would be to *test the sensitivity of the data under perturbation*. Under appropriate perturbation models, an unbiased data should give approximately the same quality of results, whereas biased data would clearly show different results under perturbation

Anonymized Data. Due to privacy concerns, data is often shared in an anonymized form. A typical form of anonymization is to ensure that each node in the network shares some neighbors with k-other nodes. An important would be to study how this *change affects the network analysis results, and to design anonymization techniques that can adapt* to the type of analysis being performed, without sacrificing the security.

To summarize, the quality and correctness of data, is an important concern for all learning algorithms. To date, there are not any standardized measures on how to evaluate quality of data or how to measure the sensitivity of results with respect to change in data. Moreover, most anonymization techniques are very general and not tuned to the analysis needs. These questions are critical to understand the limits of AI when developing smart cyberinfrastructure and to design techniques to improve the data quality.