Algorithms for analyzing evolving networks on the Dark Web & in science

Boekhout, H.D.

Summary

Large-scale data collection and analysis has become increasingly common in modern society. Whereas traditional data analysis considers information on the level of individual objects or entities in a system, network analysis studies the interaction between these objects/entities. Network analysis models the individual objects or entities as *nodes* and connects them with *edges* when an interaction occurs between them. In this thesis, we perform network analysis on two "real-world" systems while incorporating temporal information, such as the specific time of an interaction. Incorporating such temporal information can lead to richer insights. Unfortunately, there is no single standard method of incorporating temporal data that fits every analysis and research question. Consequently, for every analysis that involves temporal information, it must be considered which method of incorporation best fits both reality and the research question under consideration. Furthermore, algorithms to efficiently perform the required analysis do not always exist yet. Therefore, this thesis also introduces new efficient temporal algorithms to support our analysis.

This thesis first studies the Dark Web cryptomarket `Evolution`. This cryptomarket was active from January 2014 until March 2015 and consisted of a forum and marketplace where illegal goods and services (e.g., drugs, IDs, credit- and debitcard information, or weapons) could be traded and discussed anonymously. Chapter 2 studies the raw data on this cryptomarket that was published in the *Dark Net Market archives*. In this chapter, data quality issues are identified and resolved, the amount of missing data is estimated, and the forum and market data are linked. Data inconsistencies and data currency issues were often easily resolved or understood as actual changes when surrounding data records were complete, while duplicate data actually

served to bolster otherwise incomplete records. Consequently, it could be beneficial to intentionally collect duplicate data, i.e., duplicate webpages, when "scraping" data.

In Chapter 3, the cleaned cryptomarket dataset is used to study whether communication on the forum can be used to predict the sales success of vendors on the market. A comparison is made between traditional data analysis methods, e.g., so-called activity indicators, and network analysis methods, e.g., centrality measures. Using temporal information, both on when posts were placed and on changes in the number of sales by vendors over time, we are able to consider the predictive value for not just current, but also for future sales success. Additionally, connections between users in the communication network are restricted based on the time between their respective posts. Thus, temporal information is used to both support the research question and to improve the network's ability to model the underlying reality. Simultaneously, this approach allows the analysis to rely on purely static algorithms. The best indicator of current and future sales success was shown to be *topic engagement*, a measure of the total number of responses to topics started by a user. Meanwhile, betweenness centrality, a network centrality measure, was shown to be a better predictor of successful vendors when they are relatively inactive on the forum. Consequently, betweenness centrality is able to compensate for a weakness of topic engagement. Additionally, high betweenness centrality was shown to be a good indicator of users whose roles on the forum are vital to its function, such as administrators.

The second system studied in this thesis, is the scientific publication system. Specifically, Chapter 5 studies the evolution of the co-authorship network between scientists from different cities, as well as how this evolution differs from simulations. Based on the observation that scientific institutes tend to be closely clustered within cities, the new spatial unit of *scientific cities* is introduced to capture these clusters. When we assume that it is equally difficult to engage in face-to-face collaboration for each city, this new spatial unit allows for a fairer comparison between cities than traditional definitions of a city. After all, for the traditional definition of a city, the travel time between two institutes in a major city like New York can be vastly different from those in a relatively small city like Leiden. Collaboration between scientific cities can thus be viewed as collaborations at a distance where digital collaboration becomes more likely.

Through the use of scientific cities, Chapter 5 thus studies the network of distant scientific collaborations. The evolution of this network is studied by comparing consecutive "snapshots" of the network, where each snapshot captures a slice of the network's full timespan. Comparing the snapshots based on multiple network centrality measures, we found that the collaboration network has grown more stable over time. To compare the network evolution with simulated data, we introduced a new Evolving Degree Respecting Rewired (EDRR) network model. While synthetic network data is traditionally constructed from the ground up, EDRR networks are generated

using “real-world” network data (from the preceding snapshot) as their initial network configuration. Subsequently, the same number of edge mutations as occurred between consecutive “real-world” snapshots, for both edge additions and deletions, are applied to random node pairs in the initial network to construct the final EDRR networks. Comparison with EDRR networks suggests that new collaborations between scientific cities are more likely to occur between those that are already close within the periphery of the network.

Next, Chapter 6 studies persistent scientific teams, i.e., teams whose members frequently co-author publications. These persistent teams are determined using temporal cliques, i.e., groups of fully connected nodes, in the scientific co-authorship network. To accomplish this, Chapter 4 introduces a new algorithm to efficiently enumerate temporal cliques, significantly outperforming existing algorithms. Along with other algorithmic improvements, this performance gain was achieved by efficient pruning of the search tree. To evaluate the algorithm performance, Chapter 4 also introduces a new method of generating synthetic temporal networks of varying structural and temporal densities.

Chapter 6 shows that persistent teams are involved in the majority of scientific publications and are overrepresented on highly cited publications. Furthermore, these teams tend to experience success early in their collaboration. After the first two years, the odds of producing highly cited publications drop throughout the remainder of a team’s lifespan. This contradicts the general understanding that, as teams improve their inter-group dynamics, they become more successful over time. Notably, when persistent teams form while including members with preceding persistent collaborative experience, they are more likely to produce highly cited work early. Thus, while developing team dynamics through persistence remains valuable, success often requires a fresh impulse. This is underscored by the observation that, when team members engage in new persistent collaborations with scholars outside the team, the team’s odds of (continued) success improve. Consequently, persistent collaborations should not be discouraged, so long as they do not interfere with engaging in fresh collaborations.

A central theme of this thesis has been the correct application of temporal information. In this thesis we have shown two ways to approach this: (1) using temporal data to model the system as multiple snapshot networks, each with their own time-frame, on which static methods can be applied (Chapters 3 and 5); and (2) incorporating the temporal information directly into the network model and applying temporal methods that use this information (Chapters 4 and 6). Thus, the approach to using temporal information is more dependent on the research question being considered, than the underlying system. Additionally, models to generate synthetic temporal data to compare against “real-world” data, often do not yet exist. However, the development of these models can often build upon existing static models.