# Chapter 78
# Emergent Data Mining Tools for Social Network Analysis

**Dhiraj Murthy**
*Bowdoin College, USA*

**Alexander Gross**
*Bowdoin College, USA*

**Alex Takata**
*Bowdoin College, USA*

## ABSTRACT

*This chapter identifies a number of the most common data mining toolkits and evaluates their utility in the extraction of data from heterogeneous online social networks. It introduces not only the complexities of scraping data from the diverse forms of data manifested in these sources, but also critically evaluates currently available tools. This analysis is followed by a presentation and discussion on the development of a hybrid system, which builds upon the work of the open-source Web-Harvest framework, for the collection of information from online social networks. This tool, VoyeurServer, attempts to address the weaknesses of tools identified in earlier sections, as well as prototype the implementation of key functionalities thought to be missing from commonly available data extraction toolkits. The authors conclude the chapter with a case study and subsequent evaluation of the VoyeurServer system itself. This evaluation presents future directions, remaining challenges, and additional extensions thought to be important to the effective development of data mining tools for the study of online social networks.*

## INTRODUCTION

With the increased pervasiveness of the internet, society has seen exponential growth in digital data that has been made available on global public networks. With this rise of 'Big Data,' researchers have seen the need to identify, organize, collect, and extract this information back out of the system and into useful forms (Hammer, Garcia-Molina, Cho, Aranha, & Crespo, 1997). The fields of data mining and web-content extraction are critical to this process and have remained active areas of research, as the types and forms of data available on the Web have continued to grow and evolve. The continued growth of information on the Web - due in part to more recent trends of fully online, social,

DOI: 10.4018/978-1-4666-7230-7.ch078

and context aware computing - have made more types of data available, which are of potential use in a highly interdisciplinary range of fields. Many disciplines are looking at 'Big Data' and ways to mine and analyze these data as the key to solving everything from technical problems to better understanding social interactions. For example, large sets of tweets mined from Twitter have been analyzed to detect natural disasters (Doan, Vo, & Collier, 2011; Hughes, Palen, Sutton, Liu, & Vieweg, 2008; Murthy & Longwell, in press), predict the stock market (Bollen & Mao, 2011), and track the time of our daily rituals (Golder & Macy, 2011). As our use of blogs, social networks, and social media continues to increase, so does our creation of more web-based hyperlinked data. The successful extraction of this web-based data is of considerable research and commercial value.

Data mining often goes beyond simple information retrieval and has moved towards a meta-discovery of structures and entities hidden in seas of data. As our social interactions become increasingly mediated by Internet-based technologies, the potential to use web-based data for understanding social structures and interactions will continue to increase.

Online social networks are defined as 'web-based services that allow individuals to:

1. Construct a public or semi-public profile within a bounded system,
2. Articulate a list of other users with whom they share a connection, and
3. View and traverse their list of connections and those made by others within the system' (Boyd & Ellison, 2008).

Individuals interact within online social networks through portals such as Facebook, which create social experiences for the user by creating a personalized environment and interaction space by combining knowledge of one users' online activity and relationships with information about other networked individuals. It is through data mining algorithms that Twitter, for example, determines recommendations for users to follow or topics that may be of potential interest. One way to study social networks is by examining relationships between users and the attributes of these relationships. However, data on a blog, Facebook, or Twitter is not directly translatable into network-based data that would be useful within research praxis, and this is where the ability to perform effective data mining becomes important. Social networks typically only provide individual portal access to one's egocentric network. Put in the language of social network analysis (SNA), the visible network is constructed in relation to ego (the individual being studied) and relations of ego, known as 'alters,' are seen (e.g. Facebook friends). However, in a restricted profile environment, the alters' relationships are not revealed. In order to understand network structure (which is key to a systems perspective), the researcher must use methods like data mining in order to gather information about all users and interactions by iterating over the data. A variety of different types of tools have been developed to collect this web-based information. These tools were created for a wide array of purposes. The majority of these tools have been commercially released. Some of these tools can be used to construct profiles of individuals based on data from multiple sources. Given issues of privacy, ethical uses of these tools should be strictly employed (Van Wel & Royakkers, 2004).

Despite the existence of a variety of tools, their ease-of-use and robustness can vary widely. There are many types of networks and online communities that could qualify as a subject of network-based research. Many of these virtual organizations and networks often share key elements and structures that are common across online social networks. These could include users, groups, communications, and relationship networks between these entities. Also, unlike the