# MINING GRAPH DATA

EDITED BY


Diane J. Cook
School of Electrical Engineering and Computer Science
Washington State University
Pullman, Washington

Lawrence B. Holder
School of Electrical Engineering and Computer Science
Washington State University
Pullman, Washington




WILEY-INTERSCIENCE

A JOHN WILEY & SONS, INC., PUBLICATION

# MINING GRAPH DATA


## THE WILEY BICENTENNIAL–KNOWLEDGE FOR GENERATIONS

Each generation has its unique needs and aspirations. When Charles Wiley first opened his small printing shop in lower Manhattan in 1807, it was a generation of boundless potential searching for an identity. And we were there, helping to define a new American literary tradition. Over half a century later, in the midst of the Second Industrial Revolution, it was a generation focused on building the future. Once again, we were there, supplying the critical scientific, technical, and engineering knowledge that helped frame the world. Throughout the 20th Century, and into the new millennium, nations began to reach out beyond their own borders and a new international community was born. Wiley was there, expanding its operations around the world to enable a global exchange of ideas, opinions, and know-how.

For 200 years, Wiley has been an integral part of each generation's journey, enabling the flow of information and understanding necessary to meet their needs and fulfill their aspirations. Today, bold new technologies are changing the way we live and learn. Wiley will be there, providing you the must-have knowledge you need to imagine new worlds, new possibilities, and new opportunities.

Generations come and go, but you can always count on Wiley to provide you the knowledge you need, when and where you need it!

William J. Pesce

WILLIAM J. PESCE
PRESIDENT AND CHIEF EXECUTIVE OFFICER

Peter Booth Wiley

PETER BOOTH WILEY
CHAIRMAN OF THE BOARD

# MINING GRAPH DATA

EDITED BY


Diane J. Cook
School of Electrical Engineering and Computer Science
Washington State University
Pullman, Washington

Lawrence B. Holder
School of Electrical Engineering and Computer Science
Washington State University
Pullman, Washington




WILEY-INTERSCIENCE

A JOHN WILEY & SONS, INC., PUBLICATION

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.



***Library of Congress Cataloging-in-Publication Data***

Mining graph data / edited by Diane J. Cook, Lawrence B. Holder.
p. cm.
Includes index.
ISBN-13 978-0-471-73190-0
ISBN-10 0-471-73190-0 (cloth)
1. Data mining. 2. Data structures (Computer science) 3. Graphic methods.
I. Cook, Diane J., 1963- II. Holder, Lawrence B., 1964-
QA76.9.D343M52 2006
005.74—dc22

2006012632

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

*To Abby and Ryan, with our love.*

# CONTENTS

# PREFACE

**Data mining**, or **knowledge discovery in databases**, is a large area of study and is populated with numerous theoretical and practical textbooks. In this book, we take a focused and comprehensive look at one topic within this field: *mining data that is represented as a graph*. We attempt to cover the full breadth of the topic, including graph manipulation, visualization, and representation, mining techniques for graph data, and application of these ideas to problems of current interest.

The book is divided into three parts. Part I, Graphs, offers an introduction to basic graph terminology and techniques. In Part II, Mining Techniques, we take a detailed look at computational techniques for extracting patterns from graph data. These techniques provide an overview of the state of the art in frequent substructure mining, link analysis, graph kernels, and graph grammars. Part III, Applications, describes application of mining techniques to four graph-based application domains: chemical graphs, bioinformatics data, Web graphs, and social networks.

The book is targeted toward graduate students, faculty, and researchers from industry and academia who have some familiarity with basic computer science and data mining concepts. The book is designed so that individuals with no background in analyzing graph data can learn how to represent the data as graphs, extract patterns or concepts from the data, and see how researchers apply the methodologies to real datasets.

For those readers who would like to experiment with the techniques found in this book or test their own ideas on graph data, we have set up a Web page for the book at http://www.eecs.wsu.edu.mgd. This site contains additional information on current techniques for mining graph data. Links are also given to implementations of the techniques described in this book, as well as graph datasets that can be used for testing new or existing algorithms.

With the advent of and continued prospect for large databases containing relational and graphical information, the discovery of knowledge in such data is an important challenge to the scientific and industrial communities. Fielded applications for mining graph data from real-world domains has the potential to make significant contributions of new knowledge. We hope that this book accelerates progress toward meeting this challenge.

# ACKNOWLEDGMENTS

We would like to acknowledge and thank the many people who contributed to this book. All of the authors were very willing to help and contributed excellent material to the book. The creation of this book also initiated collaborations that will continue to further the state of the art in mining graph data. We would also like to thank Whitney Lesch and Paul Petralia at Wiley for their assistance in assembling the book and to thank the faculty and staff at the University of Texas at Arlington and at Washington State University for their continued encouragement and support of our work. Finally, we would like to thank our children, Abby and Ryan, for the joy they bring to our lives and for forcing us to talk about topics other than graphs at home.

# CONTRIBUTORS

**Indrajit Bhattacharya** University of Maryland
College Park, Maryland

**Horst Bunke** Institute of Computer Science and Applied Mathematics
University of Bern
Bern, Switzerland

**Deepayan Chakrabarti** Yahoo! Research
Sunnyvale, California

**Diane J. Cook** School of Electrical Engineering and Computer Science
Washington State University
Pullman, Washington

**Walter Didimo** Dipartimento di Ingegneria Elettronica e dell'Informazione
Università degli Studi di Perugia
Perugia, Italy

**Christos Faloutsos** School of Computer Science
Carnegie Mellon University
Pittsburgh, Pennsylvania

**Thomas Gärtner** Fraunhofer AIS
Schloß Birlinghoven
Sankt Augustin, Germany

**Lise Getoor** University of Maryland
College Park, Maryland

**David Gibson** IBM Almaden Research Center
San Jose, California

**Seth A. Grennblatt** 21st Century Technologies, Inc.
Austin, Texas

**Jiawei Han** University of Illinois at Urbana-Champaign
Urbana-Champaign, Illinois

**Lawrence B. Holder** School of Electrical Engineering and Computer Science
Washington State University
Pullman, Washington

**Tamás Horváth** Fraunhofer AIS
Schloß Birlinghoven
Sankt Augustin, Germany

**Takahiko Ito** NARA Institute of Science and Technology
Ikoma, Nara, Japan

**Istvan Jonyer** Department of Computer Science
Oklahoma State University
Stillwater, Oklahoma

**George Karypis** Department of Computer Science & Engineering
University of Minnesota
Minneapolis, Minnesota

**Nikhil Ketkar** School of Electrical Engineering and Computer Science
Washington State University
Pullman, Washington

**Ravi Kumar** Yahoo! Research, Inc.
Santa Clara, California

**Michihiro Kuramochi** Department of Computer Science & Engineering
University of Minnesota
Minneapolis, Minnesota

**Quoc V. Le** Statistical Machine Learning Program
NICTA and ANU Canberra
Canberra, Australia

**Giuseppe Liotta** Dipartimento di Ingegneria Elettronica e dell'Informazione
Università degli Studi di Perugia
Perugia, Italy

**Michel Liquière** LIRMM
Montpellier, France

**Sherry E. Marcus** 21st Century Technologies, Inc.
Austin, Texas

**Kevin S. McCurley** Google, Inc.
Mountain View, California

**Akira Mogi** Institute of Scientific and Industrial Research
Osaka University
Osaka, Japan

**Hiroshi Motoda** Institute of Scientific and Industrial Research
Osaka University
Osaka, Japan

**Melanie Moy** 21st Century Technologies, Inc.
Austin, Texas

**Michel Neuhaus** Institute of Computer Science and Applied Mathematics
University of Bern
Bern, Switzerland

**Phu Chien Nguyen** Institute of Scientific and Industrial Research
Osaka University
Osaka, Japan

**Kouzou Ohara** Institute of Scientific and Industrial Research
Osaka University
Osaka, Japan

**Takashi Okada** Department of Informatics
School of Science & Engineering
Kwansei Gakuin University
Sanda, Japan

**Masashi Shimbo** NARA Institute of Science and Technology
Ikoma, Nara, Japan

**Alex J. Smola** Statistical Machine Learning Program
NICTA and ANU Canberra
Canberra, Australia

**Andrew Tomkins** Google, Inc.
Santa Clara, California

**Takashi Washio** Institute of Scientific and Industrial Research
Osaka University
Osaka, Japan

**Stefan Wrobel** Fraunhofer AIS
Schloß Birlinghoven
Sankt Augustin, Germany
and
Department of Computer Science III
University of Bonn,
Bonn Germany

**Xifeng Yan** Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana-Champaign, Illinois

**Mohammed Zaki** Department of Computer Science
Rensselaer Polytechnic Institute
Troy, New York

# 1

# INTRODUCTION

LAWRENCE B. HOLDER AND DIANE J. COOK

*School of Electrical Engineering and Computer Science*
*Washington State University, Pullman, Washington*

The ability to mine data to extract useful knowledge has become one of the most important challenges in government, industry, and scientific communities. Much success has been achieved when the data to be mined represents a set of independent entities and their attributes, for example, customer transactions. However, in most domains, there is interesting knowledge to be mined from the relationships between entities. This relational knowledge may take many forms from periodic patterns of transactions to complicated structural patterns of interrelated transactions. Extracting such knowledge requires the data to be represented in a form that not only captures the relational information but supports efficient and effective mining of this data and comprehensibility of the resulting knowledge. Relational databases and first-order logic are two popular representations for relational data, but neither has sufficiently supported the data mining process.

The graph representation, that is, a collection of nodes and links between nodes, does support all aspects of the relational data mining process. As one of the most general forms of data representation, the graph easily represents entities, their attributes, and their relationships to other entities. Section 1.2 describes several diverse domains and how graphs can be used to represent the domain. Because one entity can be arbitrarily related to other entities, relational databases and logic have difficulty organizing the data to support efficient traversal of the relational links.

*Mining Graph Data*, Edited by Diane J. Cook and Lawrence B. Holder

Graph representations typically store each entity's relations with the entity. Finally, relational database and logic representations do not support direct visualization of data and knowledge. In fact, relational information stored in this way is typically converted to a graph form for visualization. Using a graph for representing the data and the mined knowledge supports direct visualization and increased comprehensibility of the knowledge. Therefore, mining graph data is one of the most promising approaches to extracting knowledge from relational data.

These factors have not gone unnoticed in the data mining research community. Over the past few years research on mining graph data has steadily increased. A brief survey of the major data mining conferences, such as the Conference on Knowledge Discovery and Data Mining (KDD), the SIAM Conference on Data Mining, and the IEEE Conference on Data Mining, has shown that the number of papers related to mining graph data has grown from 0 in the late 1990s to 40 in 2005. In addition, several annual workshops have been organized around this theme, including the KDD workshop on Link Analysis and Group Detection, the KDD workshop on Multi-Relational Data Mining, and the European Workshop on Mining Graphs, Trees and Sequences. This increasing focus has clearly indicated the importance of research on mining graph data.

Given the importance of the problem and the increased research activity in the field, a collection of representative work on mining graph data was needed to provide a single reference to this work and some organization and cross fertilization to the various topics within the field. In the remainder of this introduction we first provide some terminology from the field of mining graph data. We then discuss some of the representational issues by looking at actual representations in several important domains. Finally, we provide an overview of the remaining chapters in the book.

## 1.1 TERMINOLOGY

*Data mining* is the extraction of novel and useful knowledge from data. A *graph* is a set of nodes and links (or vertices and edges), where the nodes and/or links can have arbitrary labels, and the links can be directed or undirected (implying an ordered or unordered relation). Therefore, *mining graph data*, sometimes called *graph-based data mining*, is the extraction of novel and useful knowledge from a graph representation of data. In general, the data can take many forms from a single, time-varying real number to a complex interconnection of entities and relationships. While graphs can represent this entire spectrum of data, they are typically used only when relationships are crucial to the domain. The most natural form of knowledge that can be extracted from graphs is also a graph. Therefore, the *knowledge*, sometimes referred to as *patterns*, mined from the data are typically expressed as graphs, which may be subgraphs of the graphical data, or more abstract expressions of the trends reflected in the data. Chapter 2 provides more precise definitions of graphs and the typical operations performed by graph-based data mining algorithms.

While data mining has become somewhat synonymous with finding frequent patterns in transactional data, the more general term of *knowledge discovery* encompasses this and other tasks as well. *Discovery* or *unsupervised learning* includes not only the task of finding patterns in a set of transactions but also the task of finding possibly overlapping patterns in one large graph. Discovery also encompasses the task of *clustering*, which attempts to describe all the data by identifying categories or clusters sharing common patterns of attributes and relationships. Clustering can also extract relationships between clusters, resulting in a hierarchical or taxonomic organization over the clusters found in the data. In contrast, *supervised learning* is the task of extracting patterns that distinguish one set of graphs from another. These sets are typically called the positive examples and negative examples. These sets of examples can contain several graph transactions or one large graph. The objective is to find a graphical pattern that appears often in the positive examples but not in the negative examples. Such a pattern can be used to predict the class (positive or negative) of new examples. The last graph mining task is the visualization of the discovered knowledge. *Graph visualization* is the rendering of the nodes, links, and labels of a graph in a way that promotes easier understanding by humans of the concepts represented by the graph.

All of the above graph mining tasks are described within the chapters of this book, and we provide an overview of the chapters in Section 1.3. However, an additional motivation for the work in this book is the important application domains and how their data is represented as a graph to support mining. In the next section we describe three domains whose data is naturally represented as a graph and in which graph mining has been successful.

## 1.2 GRAPH DATABASES

Three domains that epitomize the tasks of mining graph data are the Internet Movie Database, the Mutagenesis dataset, and the World Wide Web. We describe several graph representations for the data in these domains and survey work on mining graph data in these domains. These databases may also serve as a benchmark set of problems for comparing and contrasting different graph-based data mining methods.

### 1.2.1 The Internet Movie Database

The Internet Movie Database (IMDb) [41] maintains a large database of movie and television information. The information is freely available through online queries, and the database can also be downloaded for in-depth analysis. This database emerged from newsgroups in the early 1990s, such as rec.arts.movies, and has now become a commercial entity that serves approximately 65 million accesses each month.

Currently, the IMDb has information on 468,305 titles and 1,868,610 people in the business. The database includes filmographies for actors, directors, writers, composers, producers, and editors as well as movie information such as titles, release

dates, production companies and countries, plot summaries, reviews and ratings, alternative names, genres, and awards.

Given such filmography information, a number of mining tasks can be performed. Some of these mining tasks exploit the unstructured components of the data. For example, Chaovalit and Zhou [9] use text-based reviews to distinguish well-accepted from poorly accepted movies. Additional information can be used to provide recommendations to individuals of movies they will likely enjoy. Melville et al. [33] combine IMDb movie information (title, director, cast, genre, plot summary, keywords, user comments, reviews, awards) with movie ratings from EachMovie [14] to predict items that will be of interest to individuals. Vozalis and Margaritis [42] combine movie information, ratings from the GroupLens dataset [37], and demographic information to perform a similar recommendation task. In both of these cases, movie and user information is treated as a set of independent, unstructured attributes.

By representing movie information as a graph, relationships between movies, people, and attributes can be captured and included in the analysis. Figure 1.1(a) shows one possible representation of information related to a single movie. This hub topology represents each movie as a vertex, with links to attributes describing the movie. Similar graphs could be constructed for each person as well. With this representation, one task we can perform is to answer the following question:

What commonalities can we find among movies in the database?

Using a frequent subgraph discovery algorithm, subgraphs that appear in a large fraction of the movie graphs can be reported. These algorithms may report discoveries such as movies receiving awards often come from the same small set of studios [as shown in Fig. 1.1(b)] or certain director/composer pairs work together frequently [as shown in Fig. 1.1(c)].

By connecting people, movies, and other objects that have relationships to each other, a single connected graph can be constructed. For example, Figure 1.2 shows how different movies may have actors, directors, and studios in common. Similarly,

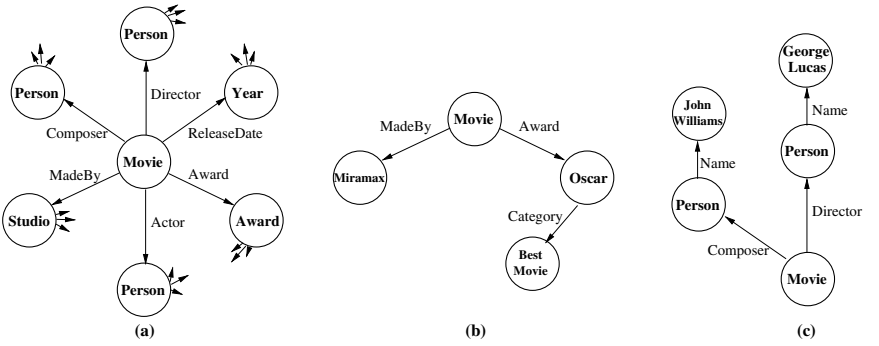


Figure 1.1. (a) Possible graph representation for information related to a single movie. (b) One possible frequent subgraph. (c) Another possible frequent subgraph.

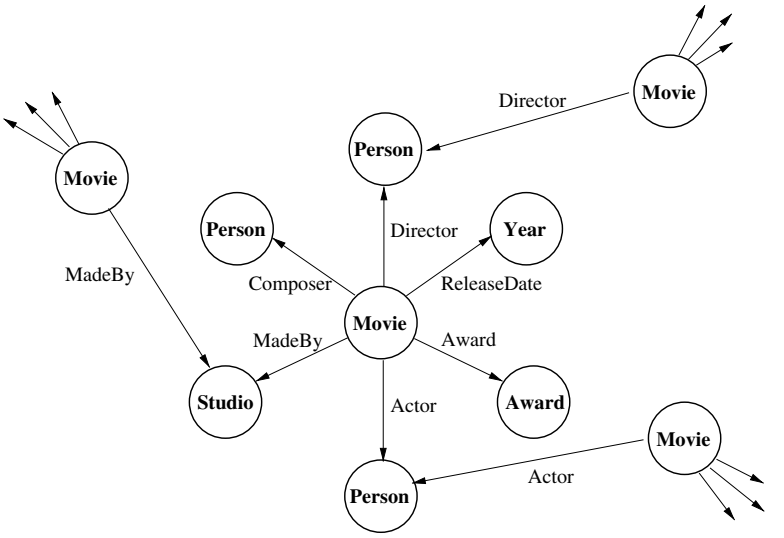


Figure 1.2. Second graph representation in which relationships between data points are represented using labeled edges.

different actors may appear in the same movie, forming a relationship between these people. Analysis of this connected graph may answer questions such as:

What common relationships can we find between objects in the database?

For the movie graph, a discovery algorithm may find a recurring pattern that movies made by the same studio frequently also have the same producer. Jensen and Neville [21] mention another type of discovery that can be made from a connected graph. In this case, an emerging film star may be characterized in the graph by a sequence of successful movies in which he or she stars and by winning one or more awards.

Other analyses can be made regarding the topology of such graphs. For example, Ravasz and Barabasi [36] analyzed a graph constructed by linking actors appearing in the same movie and found that the graph has a distinct hierarchical topology. Movie graphs can also be used to perform classification. As an example, Jensen and Neville [21] use information in a movie graph as shown in Figure 1.2 to predict whether a movie will make more than $2 million in its opening weekend. In a separate study, they use structure around nominated and nonnominated movies to predict which new movies will be nominated for awards [32].

These examples show that patterns can be learned from structural information that is explicitly provided. However, missing structure can also be inferred from this data. Getoor et al.'s [16] approach learns a graph linking actors and movies using IMDb information together with demographic information based on actor ZIP codes. Mining algorithms can be used to infer missing links in the movie graph. For example, given information about a collection of people who starred together

in a movie, link completion [17, 28] can be used to determine who the remaining individuals are who starred in the same movie. Such link completion algorithms can also be used to determine when one movie is a remake of another [21].

### 1.2.2 Mutagenesis Data

The Mutagenesis dataset is a chemical compound dataset where the task is explicitly defined as:

> Given the molecular structure of a compound, identify the compound as mutagenic or not mutagenic.

The Mutagenesis dataset was collected to identify mutagenic activity in chemical data [10]. Mutation is a structural alteration in DNA (deoxyribonucleic acid). Occasionally, a mutation improves an organism's chance of surviving or has no observable effect. In most cases, however, such DNA changes harm human health. A high correlation is also observed between mutagenicity and carcinogenicity. Some chemical compounds are known to cause frequent mutations. Mutagenicity cannot be practically determined for every compound using biological experiments, so accurate evaluation of mutagenic activity from chemical structure is very desirable.

Structure–activity relationships (SARs) relate biological activity with molecular structure. The importance of SARs to drug design is well established. The Mutagenesis problem focuses on obtaining SARs that describe the mutagenicity of nitroaromatic compounds or organic compounds composed of NO or $NO_2$ groups attached to rings of carbon atoms. Analyzing relationships between mutagenic activity and molecular structure is of great interest because highly mutagenic nitroaromatic compounds are carcinogenic.

The Mutagenesis dataset collected by Debnath et al. [10] consists of the molecular structure of 230 compounds, such as the one shown in Figure 1.3. Of these compounds, 138 are labeled as mutagenic and 92 are labeled nonmutagenic. Each compound is described by its constituent atoms, bonds, atom and bond types, and partial charges on atoms. In addition, the hydrophobicity of the compound (log $P$), the energy level of the compound's lowest unoccupied molecular orbital (LUMO), a Boolean attribute identifying compounds with three or more benzyl rings (I1), and a

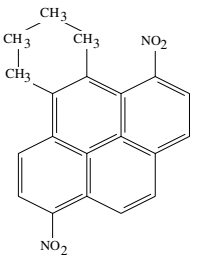


Figure 1.3. 1,6,-Dinitro-9,10,11,12-tetrahydrobenzo[$e$]pyrene.

Boolean attribute identifying compounds that are acenthryles (Ia). The mutagenicity of the compounds has been determined using the Ames test [1]. While alternative datasets are being considered by the community as challenges for structural data mining [29], the Mutagenesis dataset provides both a representative case for graph representations of chemical data and an ongoing challenge for researchers in the data mining community.

Some work has focused on analyzing these chemical compounds using global, nonstructural descriptors such as molecular weight, ionization potential, and various physiocochemical properties [2, 19]. More recently, researchers have used inductive logic programming (ILP) techniques to encode additional relational information about the compounds and to infuse the discovery process with background knowledge and high-level chemical concepts such as the definitions of methyl groups and nitro groups [23, 40]. In fact, Srinivasan and King in a separate study [38] show that traditional classification approaches such as linear regression improve dramatically in classification accuracy when enhanced with structural descriptors identified by ILP techniques.

Inductive logic programming methods face some limitations because of the explicit encoding of structural information and the prohibitive size of the search space [27]. Graphs provide a natural representation for the structural information contained in chemical compounds. A common mining task for the Mutagenesis data, therefore, is to represent each compound as a separate graph and look for frequent substructures in these graphs. Analysis of these graphs may answer the following question:

> What commonalities exist in mutagenic or non-mutagenic compounds that will help us to understand the data?

This question has been addressed by researchers with notable success [5, 20]. A related question has been addressed as well [11, 22]:

> What commonalities exist in mutagenic or nonmutagenic compounds that will help us to learn concepts to distinguish the two classes?

An interesting twist on this task has been offered by Deshpande et al. [12], who do not use the substructure discovery algorithm to perform classification but instead use frequency of discovered subgraphs in the compounds to form feature vectors that are then fed to a Support Vector Machine classifier.

Many of the graph templates used for the Mutagenesis and other chemical structure datasets employ a similar representation. Vertices correspond to atoms and edges represent bonds. The vertex label is the atom type and the edge label is the bond type. Alternatively, separate vertices can be used to represent attributes of the atoms and the bonds, as shown in Figure 1.4. In this case information about the atom's chemical element, charge, and type (whether it is part of an aromatic ring) is given along with attributes of the bond such as type (single, double, triple) and relative three-dimensional (3D) orientation. Compound attributes including log

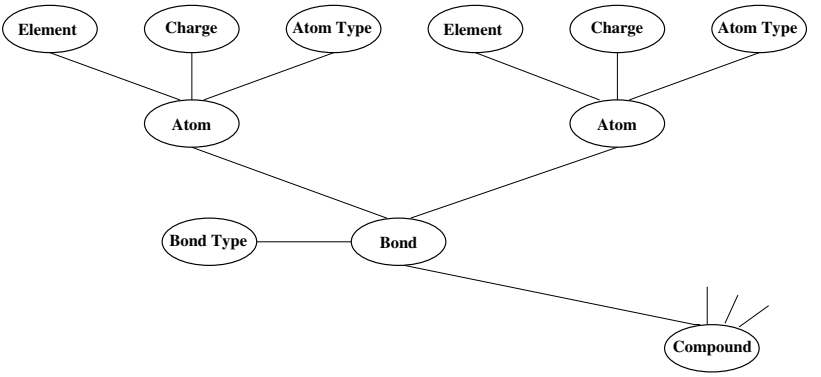


Figure 1.4. Graph representation for a chemical compound.

*P*, LUMO, I1, and Ia can be attached to the vertex representing the entire chemical compound.

When performing a more in-depth analysis of the data, researchers often augment the graph representation with additional features. The types of features that are added are reflective of the type of discoveries that are desired. Ketkar et al. [22], for example, add inequality relationships between atom charge values with the goal of identifying value ranges in the concept description. In Chapter 14 of this book, Okada provides many more descriptive features that can be considered.

### 1.2.3 Web Data

The World Wide Web is a valuable information resource that is complex, dynamically evolving, and rich in structure. Mining the Web is a research area that is almost as old as the Web itself. Although Etzioni coined the term "Web mining" [15] to refer to extracting information from Web documents and services, the types of information that can be extracted are so varied that this has been refined to three classes of mining tasks: Web content mining, Web structure mining, and Web usage mining [26].

Web content mining algorithms attempt to answer the following question:

What patterns can I find in the content of Web pages?

The most common approach to answering this question is to perform mining of the content that is found within each page on the Web. This content typically consists of text occasionally supplemented with HTML tags [8, 43]. Using text mining techniques, the discovered patterns facilitate classification of Web pages and Web querying [4, 34, 44].

When structure is added to Web data in the form of hyperlinks, analysts can then perform Web structure mining. In a Web graph, vertices represent Web pages and edges represent links between the Web pages. The vertices can optionally be