AI Searching Techniques

Measuring Similarity and Distance

Dr.Sherif Saad

Learning Objectives

- Understanding and building simple similarity measures.
- Study complex similarity measures and algorithms
- Applications of similarity measures in AI

Last-Time

Meaning of Similarity

Types of Similarity Measures

Virtual Attributes

Outlines

- The Local-Global Principle for Similarity Measures
- Graph Representations and Graph Similarities
- Taxonomic Similarities

Local-Global Principle for Similarity Measures

Each object ***A*** is constructed from atomic parts $\boldsymbol{A_i}$ that are attributes by some construction process:

$$A = C(A_i | i \in I)$$

To measure the similarity we compare first the atomic parts and then to compare more complex constructs.

The comparison of the atomic parts reflects a local view and will be done by so-called local measures. The comparison of the whole objects reflects a global view.

Local-Global Principle for Similarity Measures

This means **sim(a,b)** is of the form

$$\mathrm{sim}(a, b) = \sum_{i=1}^{i=n} \omega_i \, \mathrm{sim}_i (a_i, b_i)$$

Combine symbolic with numerical arguments. An example for this is the Heterogeneous Euclidean Overlap Metric (HEOM):

$$HEOM(x, y) = \sqrt{\sum_{i=1}^{n} d_i(x_i, y_i)^2}, \qquad d_i = \begin{cases} H(x_i, y_i), \\ \mathrm{dist}(x_i, y_i), \end{cases}$$

Local Measures

- The local measures should represent domain properties and their relation to the task in question.
- They are in general not uniform even if the domain is a numerical interval
- The domain of the problem affects the importance of the attribute.
- Local attribute can contain important knowledge for solving the problem. Because the knowledge can be of arbitrary nature, there is no general method to represent it in the measure.

Exercise: Attributes Importance

- List 5 attributes (properties) that we can use to describe the following objects: Car, Mobile Phone, Book, Soccer Team
    - Sort these attributes by importance
    - Think of two different applications that affect the importance of these attributes

Attribute Importance

We can assign weights to attributes to indicate the attribute importance

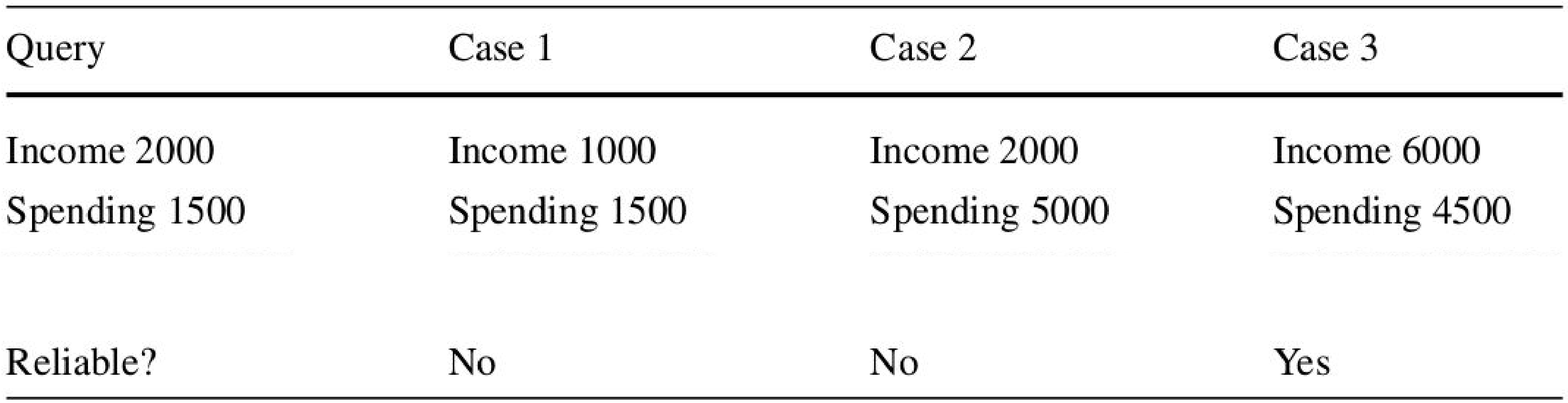
$$\text{Weighted average:} \quad F(x_1, \ldots, x_n) = \sum_{i=1}^{n} \omega_i \cdot x_i$$

$$\text{with} \sum_{i=1}^{n} \omega_i = 1 \quad \text{and} \quad x_i = \text{sim}_i(a_i, b_i),$$

Two attributes ***A*** and ***B*** are independent if their importance is not coupled, i.e., the change in the influence of one attribute on the solution has no effect on the influence of the other attribute to the solution.

Virtual Attributes

Let assume that the following table show reliability for getting a loan from a bank



Query	Case 1	Case 2	Case 3
Income 2000	Income 1000	Income 2000	Income 6000
Spending 1500	Spending 1500	Spending 5000	Spending 4500
Reliable?	No	No	Yes

Virtual Attributes (cont...)

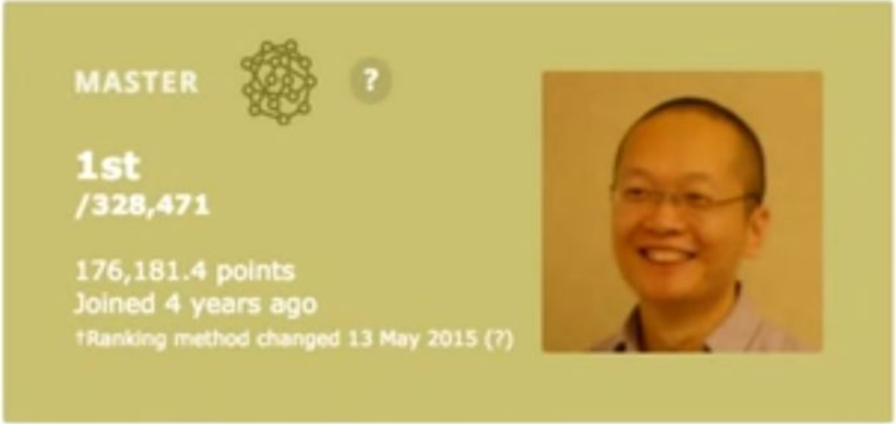
An attribute that is not explicitly define in the data (problem)

It is usually derived or defined from other attributes

A large problem is to find and define virtual attributes.

I usually use other people's code [...] it is usually not "efficient" (from time budget perspective) to write my own algorithm [...] I can find open source code for what I want to do, and my time is much better spent doing research and feature engineering -- Owen Zhang

http://blog.kaggle.com/2015/06/22/profiling-top-kagglers-owen-zhang-currently-1-in-the-world/



MASTER

1st
/328,471

176,181.4 points
Joined 4 years ago
†Ranking method changed 13 May 2015 (?)

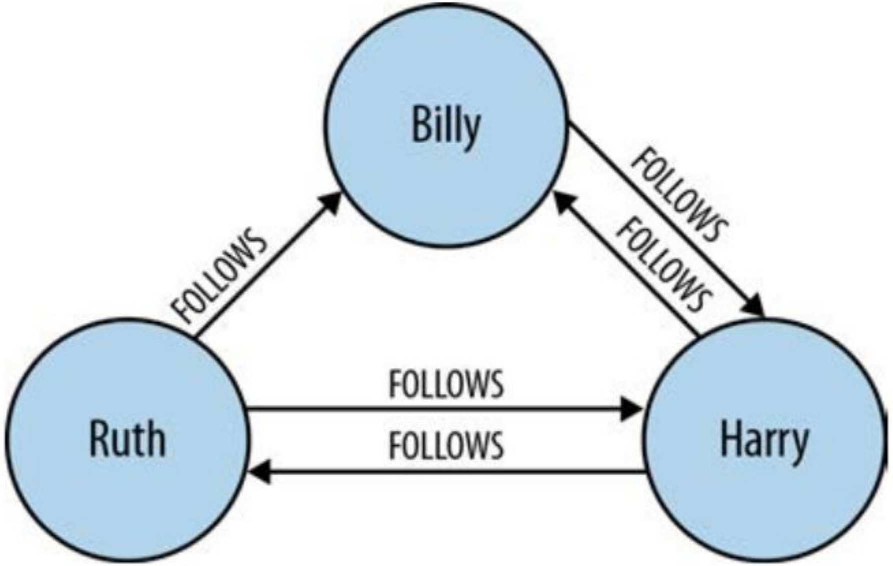
Graph Representations and Similarities

- Trees and graphs are very useful for representing certain complex objects.
- The size of a graph is usually just the number of edges; the order of a graph is the number of nodes.
- An attributed graph is a directed graph with additional marks (labels) on nodes and edges:
- Attributed graphs are often called semantic nets (property graph)

Graph Similarities: Attributed Graph

Attributed Property Graph:

- It contains nodes and relationships
- Nodes contain properties (key-value pairs)
- Relationships are named and directed, and always have a start and end node
- Relationships can also contain properties



Graph Similarities: Isomorphism

- Two graphs as similar if they are isomorphic. Isomorphism of graphs ***G*** and **H** is a **bijection** between the vertex sets of ***G*** and ***H***
- Any two vertices ***u*** and ***v*** of ***G*** are adjacent in ***G*** if and only if ***f(u)*** and ***f(v)*** are adjacent in ***H***
- They have the same number of nodes (vertices), degree, and shape, Is this means similar or equal?
- A graph G_1 is subgraph isomorphic to a graph G_2 if there is a subgraph G'_2 of G_2 which is isomorphic to G_1

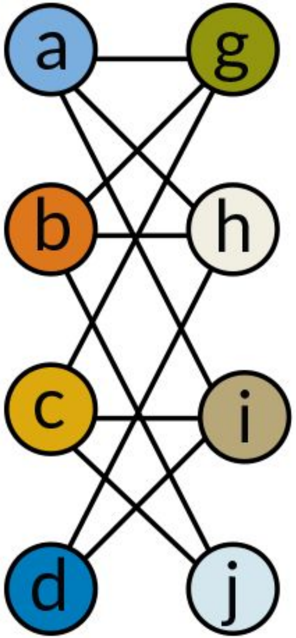
Graph Similarities: Largest Common Subgraphs

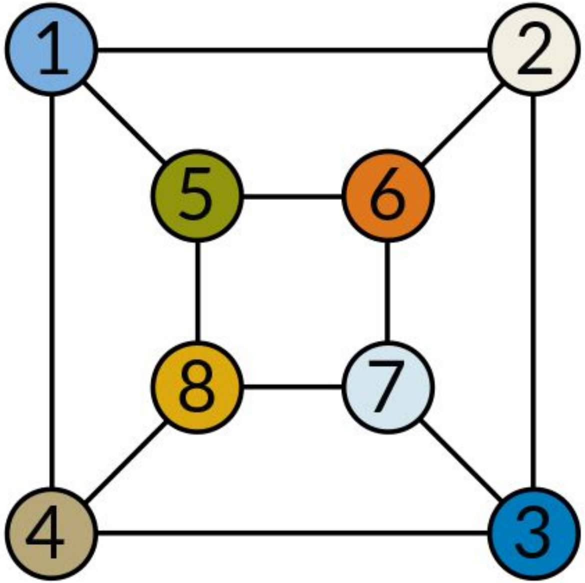
- A graph **G** is a largest common subgraph of ***G***$_1$ and ***G***$_2$ if and only if
- ***G*** is subgraph isomorphic to **G**$_1$ and subgraph isomorphic to **G**$_2$
- There is no graph ***G'*** that is subgraph isomorphic to ***G***$_1$ and ***G***$_2$ such that **|*G'*| > |G|**.
- This allows a numerical similarity measure:

$$\mathrm{sim}\, G_{graph}(G_1, G_2) = |G'| \quad \text{where } G' \text{ is a largest common subgraph of } G_1 \text{ and } G_2$$

Graph Similarities: Isomorphism

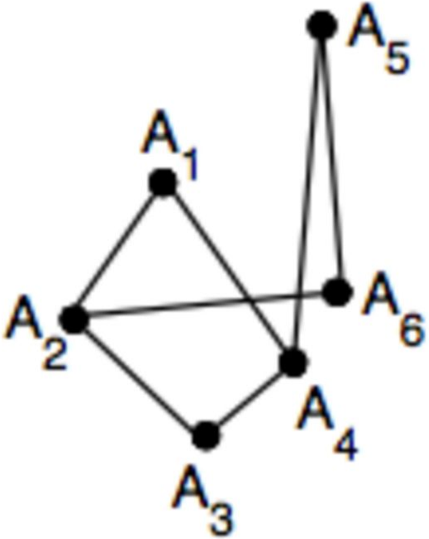
These two graphs are isomorphic

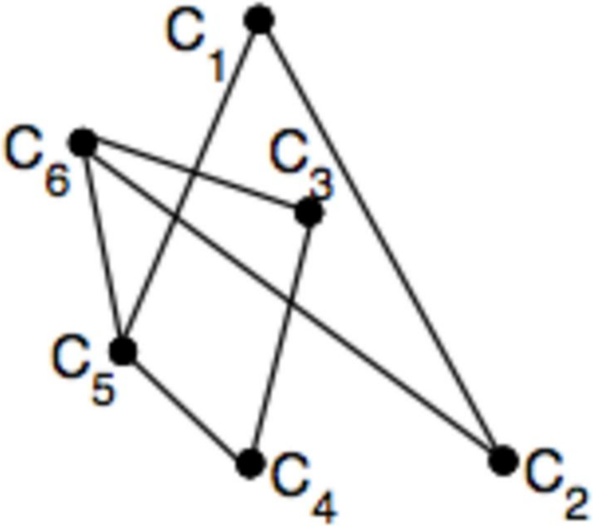




Exercise: Graph Similarities

Are these two graphs isomorphic?





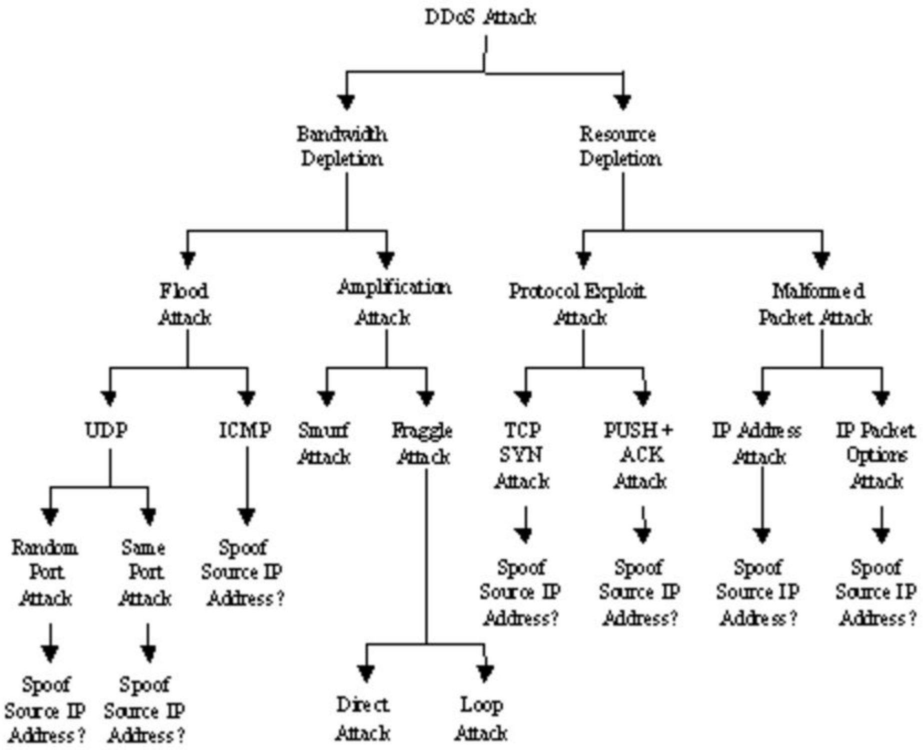
Graph Similarities: Edit Operations

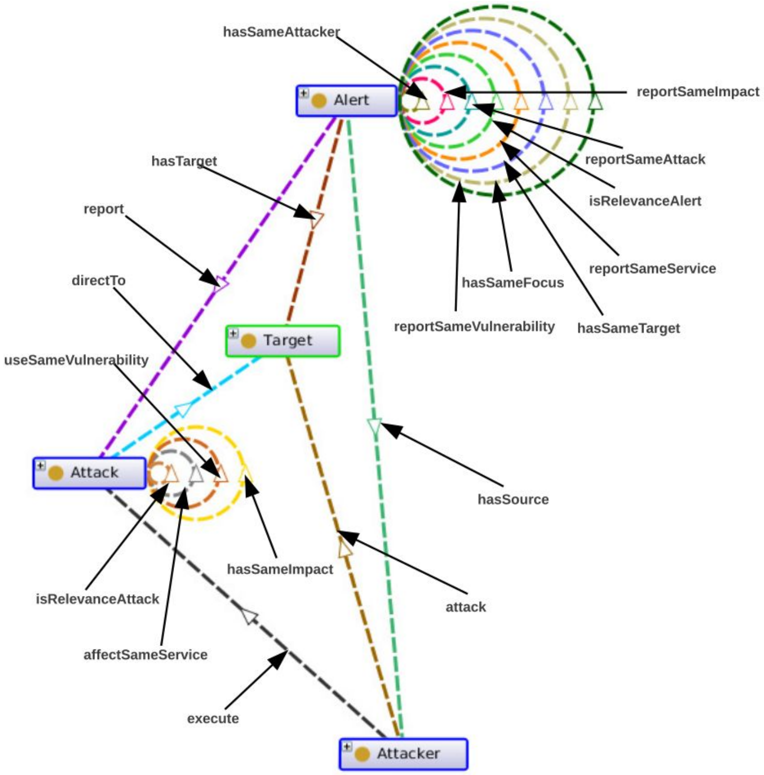
- The basic edit operations over a property graph
  - Inserting nodes.
  - Deleting nodes.
  - Inserting edges.
  - Deleting edges.
  - Replacements.
  - Changing marks on nodes and edges.
- Two graphs G and H are more similar the easier or cheaper the transformations are.
- In your opinion which edit operation is the most expensive one and why?

Taxonomic Similarities

- Taxonomies are widely used structures and specialisations of graphs
- They are efficient method to model concepts in specific domains going from general to more specific objects.
- They are the main building blocks of Ontologies and very common method in knowledge representation.
- Taxonomy reflects semantic relations, for instance, two successor nodes of a node have something in common
- One way to measure semantic similarity

Examples: Taxonomies and Ontologies





Taxonomic Similarities

- The simplest distance measure is the graph distance

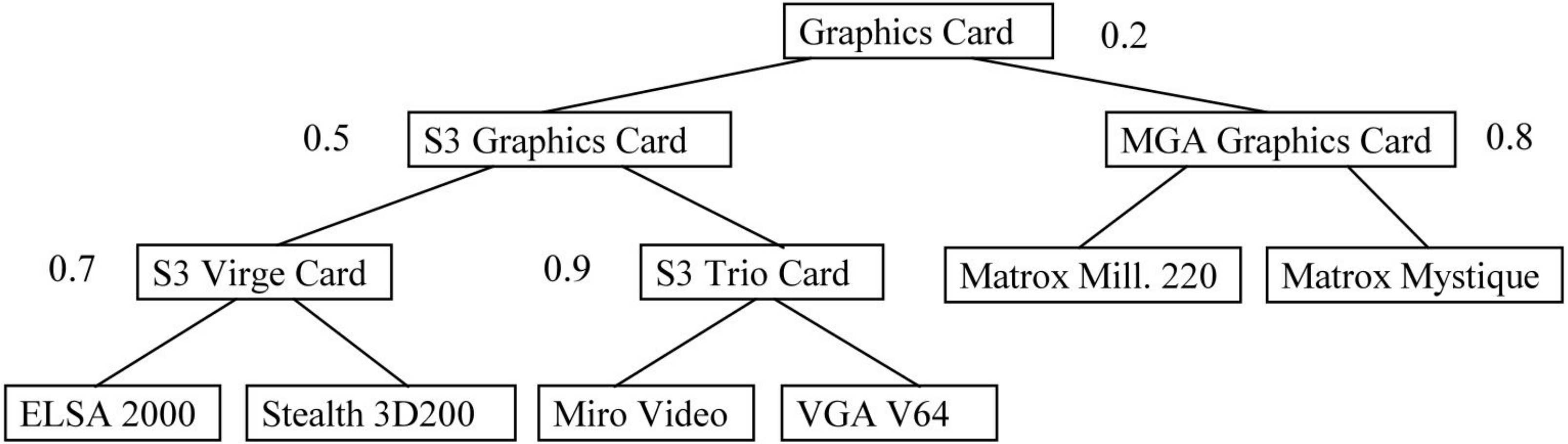
$$d_{gr}(u, v) = \text{length of shortest path from } u \text{ to } v$$

- This is the path through the deepest common predecessor of u and v. This predecessor is called DCP (Deepest Common Predecessor).
- Wu–Palmer metric. It measures the depth of two given concepts in the tax

$$\mathrm{sim}_{WuPal}(c1, c2) = \frac{2 \cdot depth(DCP(c1, c2))}{depth(c1) + depth(c2)}.$$

Exercise: Taxonomic Similarities

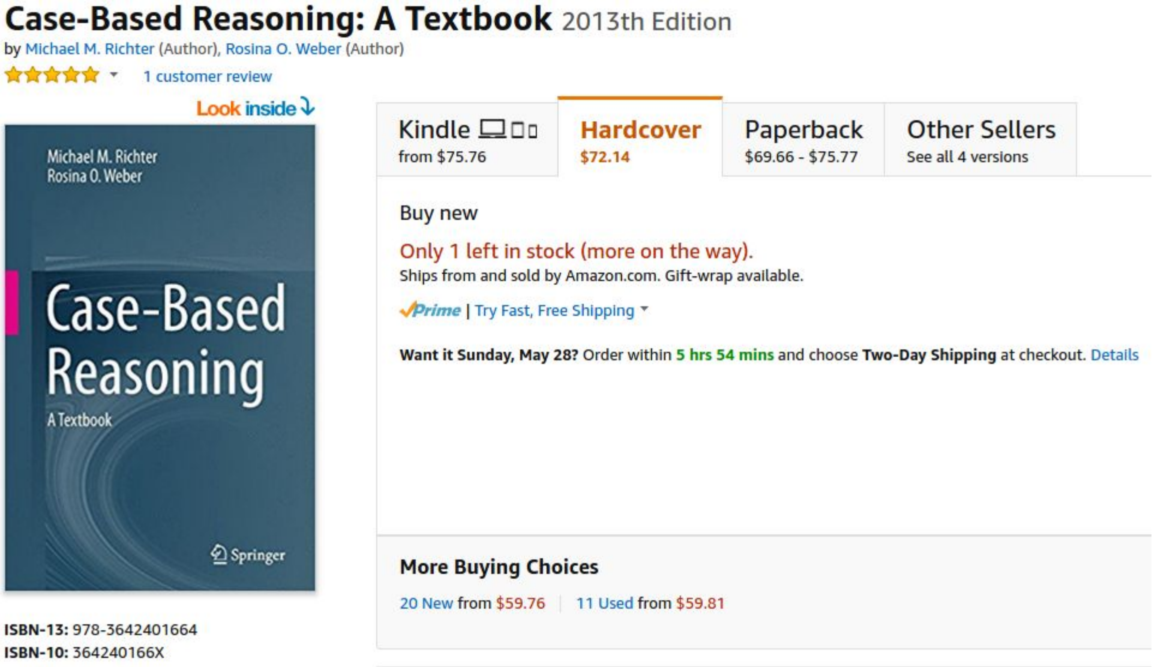
- Given the following graphic cards taxonomy, find the similarity between:
  - Miro Video and ELSA 2000
  - S3 Graphics Card and MGA Graphics Card
  - VGA V64 and S3 Trio Card



Additional Reference

The slides of today class are based on the textbook: Case-Based Reasoning: A Textbook

By Michael M. Richter



Case-Based Reasoning: A Textbook 2013th Edition

by Michael M. Richter (Author), Rosina O. Weber (Author)

★★★★★ ▾ 1 customer review

Kindle	**Hardcover**	Paperback	Other Sellers
from $75.76	$72.14	$69.66 - $75.77	See all 4 versions

Buy new

Only 1 left in stock (more on the way).

Ships from and sold by Amazon.com. Gift-wrap available.

✓Prime | Try Fast, Free Shipping ▾

Want it Sunday, May 28? Order within 5 hrs 54 mins and choose **Two-Day Shipping** at checkout. Details

More Buying Choices

20 New from $59.76 | 11 Used from $59.81

ISBN-13: 978-3642401664
ISBN-10: 364240166X

Questions