
Constructions from Dots and Lines

Marko A. Rodriguez¹ and Peter Neubauer²

¹ AT&T Interactive markorodriguez@attinteractive.com

² NeoTechnology peter.neubauer@neotechnology.com

Summary. A graph is a data structure composed of dots (i.e. vertices) and lines (i.e. edges). The dots and lines of a graph can be organized into intricate arrangements. The ability for a graph to denote objects and their relationships to one another allow for a surprisingly large number of things to be modeled as a graph. From the dependencies that link software packages to the wood beams that provide the framing to a house, most anything has a corresponding graph representation. However, just because it is possible to represent something as a graph does not necessarily mean that its graph representation will be useful. If a modeler can leverage the plethora of tools and algorithms that store and process graphs, then such a mapping is worthwhile. This article explores the world of graphs in computing and exposes situations in which graphical models are beneficial.

1 The Bits and Pieces of the Dots and Lines

A model is a representation of some aspect of reality. Many models can be thought of as a collection of objects (e.g. people, concepts) and the relationships that exist between them (e.g. friendships, subclasses). Such objects and relations form a network. Graphically, an object in a network can be denoted by a dot and a relationship can be denoted by a line. A structure formed by dots and lines is known as a graph—the mathematical term for a network [13]. The most common type of graph is the simple graph. An example instance is diagrammed in Figure 1. In a simple graph there are a set of vertices (i.e. dots) and a set of edges (i.e. lines), where edges are undirected, connect two unique vertices (i.e. no loops), and no two edges exist between the same pair of vertices.

Contrary to the title of this article, dots and lines are not the only components in a graph modeler’s toolkit. There are many more bits and pieces

¹ Rodriguez, M.A., Neubauer, P., “Constructions from Dots and Lines,” Bulletin of the American Society for Information Science and Technology, volume X, number X, pages X–X, ISSN:1550-8366, 2009.

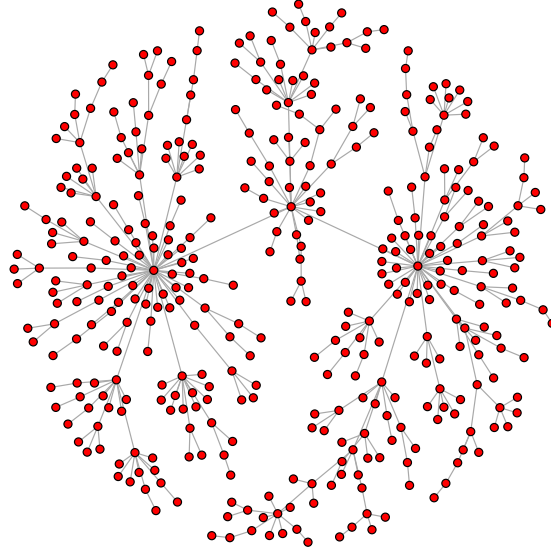


Fig. 1. The prototypical graph is the simple graph. In this structure there exists dots (i.e. vertices) and lines (i.e. edges). While the primitives are simple, their amalgamation can yield great complexity.

in the world of graphs. In practice, rarely are vertices and edges the only data contained within a graph. For instance, sometimes its useful to have a name associated with a vertex, a weight and direction associated with an edge, etc. From primitive dots and lines various bits and pieces can be added to yield a more flexible, more expressive graph. Figure 2 diagrams a collection of different graph types. A short summary of each graph type is provided in the itemization below. Note that in many cases, these bits and pieces can be used in combination with one another (i.e. they are not necessarily mutually exclusive).

- **half-edge graph:** a unary edge (i.e. an edge that “connects” one vertex) has limited practical application and is primarily discussed in mathematics.
- **multi-graph:** there are many situations in which it is desirable to have multiple edges between the same two vertices.
- **simple graph:** the prototypical graph, where an edge connects two vertices and no loops are allowed.
- **weighted graph:** used to represent strength of ties or transition probabilities.
- **vertex-labeled graph:** most every graph makes use of labeled vertices (e.g. an identifier).
- **semantic graph:** used to model cognitive structures such as the relationship between concepts and the instances of those concepts [12].

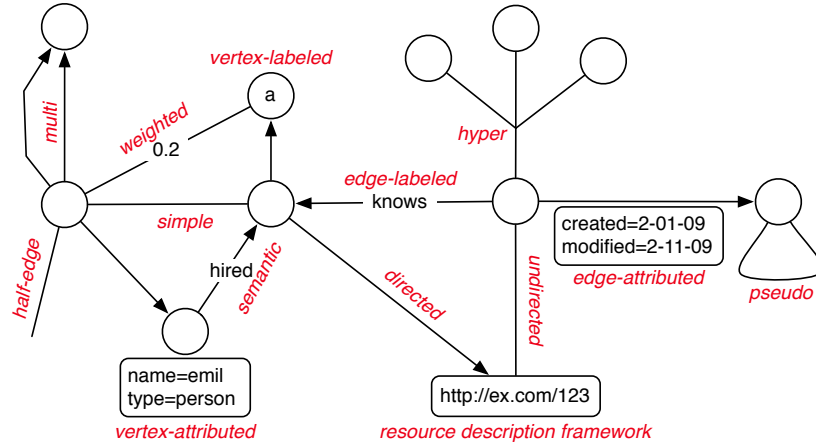
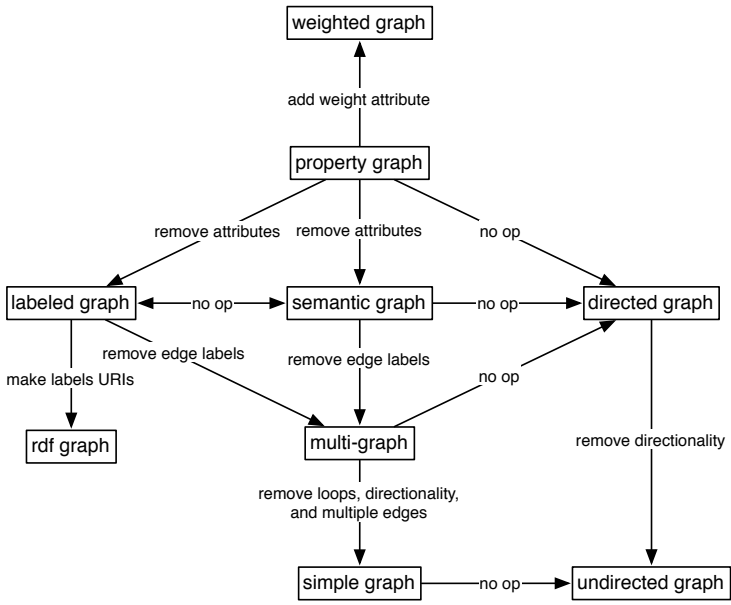


Fig. 2. There are numerous types of graphs. Many of the formalisms described can be mixed and matched in order to provide the modeler the expressivity necessary to capture the essential features of a domain.

- **vertex-attributed:** used in applications where it is desirable to append non-relational metadata to a vertex.
- **edge-labeled graph:** used to denote the way in which two vertices are related (e.g. friendships, kinships, etc.).
- **directed graph:** orders the vertices of an edge to denote edge orientation.
- **hypergraph:** generalizes a binary edge whereby an edge connects an arbitrary number of vertices [6].
- **undirected graph:** the typical graph that is used when the relationship is symmetric (e.g. friendship).
- **resource description framework graph:** a graph standard developed by the the World Wide Web consortium that denotes vertices and edges by Uniform Resource Identifiers [7].
- **edge-attributed graph:** used in applications where it is desirable to append non-relational metadata to an edge.
- **pseudo graph:** used to denote a reflexive relationship.

The list presented is not the complete space of all graph types, nor are the terms generally accepted in all domains. Many of these structures have been rediscovered in different domains and under different names. The important point is that there are numerous graph types and, consequently, there are systems and algorithms that exist to store and process them.

A common graph type supported by most graph systems is the directed, labeled, attributed, multi-graph—also known as a “property graph.” Graphs of this form allow for the representation of labeled vertices, labeled edges, and attribute metadata (i.e. properties) for both vertices and edges. The property graph is common because by simply abandoning or adding particular bits



2 Preserving Dots and Lines

The computer science community has recently seen an explosion of database technologies. For decades, the relational database of Codd’s relational algebra has been the primary storage and query mechanism for large data sets [4]. However, with the continued growth of data and an increasingly variegated application landscape, new databases have emerged. In this space, no database is seen as the single solution to all problems. Instead, each database attempts to solve a particular data management issue. Itemized below is a short description of recent database types.

- **document database:** These databases have the “document” as their atomic entity. Such objects are semi-structured and usually represented in XML or JSON. A document can be retrieved by means of pattern matching a query document (i.e. a semi-populated document) against all the documents contained in the database. The benefit of this model is that these databases scale horizontally with relative ease. This is due to the fact that documents lack references between one another. The drawback is that data is not interrelated and thus, cross database analyses are costly. For many web applications the document databases is a well-suited solution that supports data scale and a convenient symmetry between the document structure and the processing language (e.g. languages that natively support XML and/or JSON). Examples of such databases include MongoDB⁵ and CouchDB⁶.
- **key/value store:** This family of databases is focused on the scaling of large amounts of data over a large number of machines and, in turn, supporting heavy read/write loads. Most of the databases in this class were inspired by Amazon’s Dynamo [5]. A popular open-source key/value store is Tokyo Cabinet.⁷
- **triple/quad store:** Triple/quad-stores were developed to support the demands of the Semantic Web/Web of Data/Linked Data community. These databases are optimized for storing and querying data represented according to the Resource Description Framework (RDF) [7]. Typical use cases include description logic reasoning [1] and SPARQL-based graph pattern matching [8]. AllegroGraph is a high-performance quad-store with a large suite of extensions and features.⁸
- **column store:** Most column stores are modeled after Google’s BigTable database [3]. A big table is a sparse, distributed, persistent multi-dimensional sorted map. The map is indexed by a row key, column key, and a time-stamp. Real-world services implemented with BigTable include Google-

⁵ MongoDB is available at <http://www.mongodb.org/>.

⁶ CouchDB is available at <http://couchdb.apache.org/>.

⁷ Tokyo Cabinet is available at <http://1978th.net/tokycabinet/>.

⁸ AllegroGraph is available at <http://www.franz.com/agraph/allegrograph/>.

Analytics and GoogleEarth. Cassandra is a popular open-source column store.⁹

- **graph database:** Graph databases are optimized for the efficient processing of dense, interrelated datasets. In these databases, the atomic entity is the graph as a whole. The typical data model is the property graph. By supporting the interrelation of data, graph databases allow for fast traversals along the edges between vertices [9]. A popular graph database of this form is Neo4j.¹⁰

There are numerous databases in this growing space that were not mentioned. Moreover, there are other databases types not mentioned. It is out of the scope of this article to explore this space in depth. The interested reader is directed to related discussions, blog posts, and presentations that are made freely available on the Internet. Of particular relevance to this article is the graph database and the property graph data model. Figure 4 diagrams a property graph containing people, their articles, and a university. In this particular domain model, each vertex has a **name** property and a **type** property. Edges denote both a directionality and a relationship type (i.e. an edge label). Moreover, its possible to also include properties on an edge to further refine the way in which two vertices are related (e.g. Josh started attending RPI in 2007).

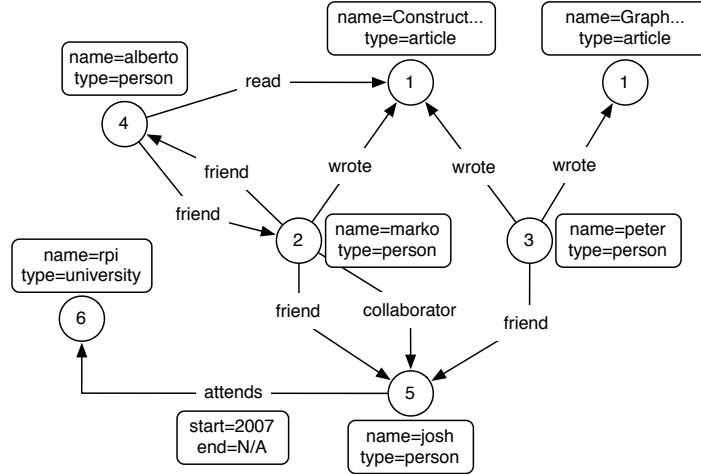


Fig. 4. A property graph is a directed, labeled, attributed, multi-graph. The edges are directed, vertices/edges are labeled, vertices/edges have associated key/value pair metadata (i.e. properties), and there can be multiple edges between any two vertices.

⁹ Cassandra is available at <http://cassandra.apache.org/>.

¹⁰ Neo4j is available at <http://neo4j.org/>.

A consequence of the flexibility of a graph is that other graph structures can be represented along with the domain model. A typical use case of such graph extensions include endogenous indices. An index is usually a tree-structure that allows for the fast look-up of elements within a collection. If there were no indices into a collection, then to determine if a particular element had a particular property, each element in the collection would have to be examined. The cost of a linear scan of this kind is $\mathcal{O}(n)$, where n is the number of elements. What an index provides is the ability to partition the elements into increasingly fine-grained bins. Most indices have a lookup cost of $\mathcal{O}(\log_2 n)$. While an index creates more data (the tree structure), it makes up for this cost by greatly increasing the speed of element retrieval. Figure 5 demonstrates a **name**-property index over the example graph diagrammed in Figure 4. Together, the domain model and the index of the domain model are seen as a single atomic entity. Searching for an element and moving between elements are accomplished by a unified framework: the graph traversal.

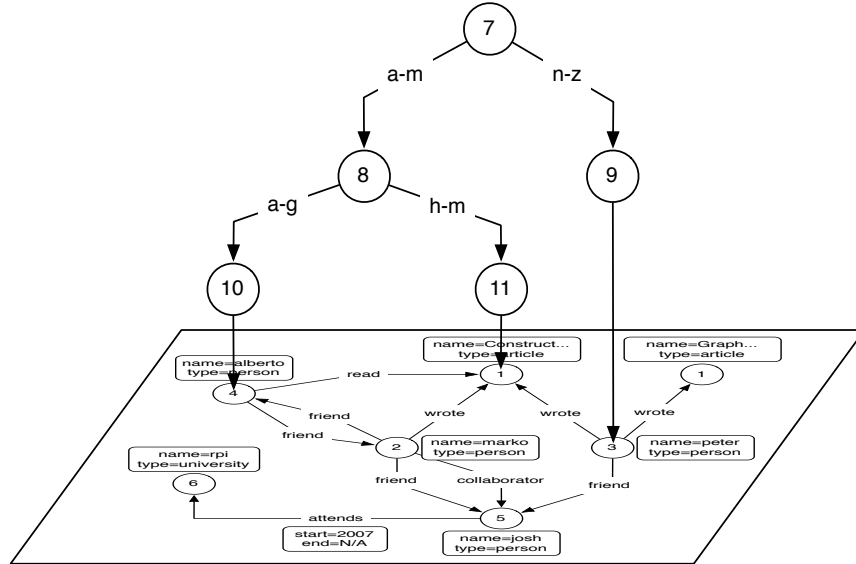


Fig. 5. The index of the attributes/properties of the vertices and edges tend to be trees. A graph is a generalization of a tree. As such, graph databases allow for the modeling of the indices of the graph within the graph structure itself. For the sake of diagram clarity, the index does not touch every vertex with a **name** property. Finally, the edge labels of the index tree denote the “bin” that each sub-vertex is representing.

3 Jumping from Dot to Dot

The first aspect of using a graph is creating a graph. Once a graph has been created, it can be subjected to algorithms that quantify aspects of its structure, alter its structure, or solve-problems that are a function of its structure. At the root of any of these algorithms is the graph traversal [9]. A graph traversal is a “walk” along the elements of a graph—from vertex, to edge, to vertex, etc. As this walk proceeds, aspects of the graph can be saved or manipulated and in general, an algorithm can be computed. In principle, any of the data models and databases presented in the previous section (and including typical relational databases) can be used to represent and process a graph. However, when traversing a graph is the ultimate use case for a graph data set, then a graph databases is the optimal solution.¹¹

To get a better understanding of how graph traversals work, the examples in this section will be expressed in terms of a graph programming language called Gremlin.¹² In Gremlin, moving over vertices and edges is analogous, in many ways, to moving through the directory structure of a local filesystem. To demonstrate, a naïve friend-of-a-friend query is represented as follows:

```
./outE[@label='friend']/inV/outE[@label='friend']/inV
```

Reading from left to right, this expression states:

- Start at the root vertex (`.`, i.e. the vertex to evaluate the expression on).
- Traverse to all the outgoing edges of the root vertex (`/outE`).
- Filter out all edges that are not labeled “friend” (`[@label='friend']`).
- For all those friend-labeled edges, go to their incoming/head vertices (`/inV`).
- For all the friends of the root vertex, get their outgoing edges (`/outE`).
- Filter out all edges that are not labeled “friend” (`[@label='friend']`).
- For all those friend-labeled edges, go to their incoming/head vertices (`/inV`).

At the end of this expression, the resultant vertices are the friends of the friends of the root vertex. Figure 6a diagrams the traversal, where the grey vertices are the returned vertices. This example is “naïve” because in many cases, its important to retrieve the root vertex’s friends of friends that are not also its friends. In such situations, the traverser must remember if a located friend-of-friend was not already a friend. In order to calculate the friend-of-a-friend, the friends must be determined first. Therefore, its possible to save this information for later use. This idea is diagrammed in Figure 6b and the

¹¹ This is an import point. A graph database is optimized for graph traversals because elements (i.e. vertices and edges) maintain direct references to their adjacent elements. It is this design choice that makes traversing a graph structure within a graph database fast and efficient.

¹² Gremlin is available at <http://gremlin.tinkerpop.com/>.

Gremlin expression is presented below, where the variable $\$x$ references the friends of the root vertex.

```
./outE[@label='friend']/inV[g:assign('$x')]/
  outE[@label='friend']/inV[g:except($x)]
```

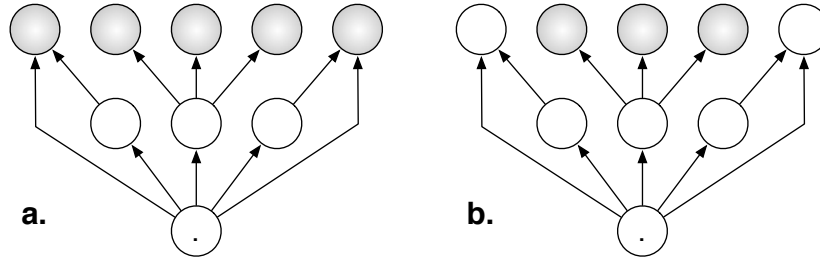


Fig. 6. a.) The grey vertices denote the friends of the friends of the root vertex. b.) The grey vertices denote the friends of the friends of the root vertex who are not also the friends of the root vertex. For the sake of diagram clarity, the edges are not labeled. Assume that all edges are labeled “friend.”

An important aspect of working with property graphs is that the edges are typed/labeled. The standard suite of graph algorithms found in most graph/network-theory textbooks are not immediately useful for property graphs [2]. This is because, most graph algorithms have been developed for unlabeled graphs. When vertices can be related by many different ways and vertices can represent various types of objects, the meaning of the rankings, paths, etc. returned by standard graph algorithms are ambiguous. However, by interpreting a path through a graph as an edge, it's possible to express standard graph algorithms on property graphs [10]. The previously presented Gremlin expression followed a path from the root vertex to its friends' friends. This path can be considered a “virtual” (i.e. inferred, derived) edge. From the perspective of this expression, a new implicit graph is created over the graph's vertices that only contains edges labeled “friend-of-a-friend.” This idea is diagrammed in Figure 7. As such, this “virtual” graph is equivalent to an unlabeled graph because all edges having the same meaning. Therefore, all the standard graph algorithms can be meaningfully applied to this derived graph—e.g. the shortest path between person *A* and person *B* through their friends of friends. The benefit of edge-labeled graphs (e.g. property graphs) is that there are as many types of rankings, scorings, etc. as there are types of paths that exist between the elements of the graph.

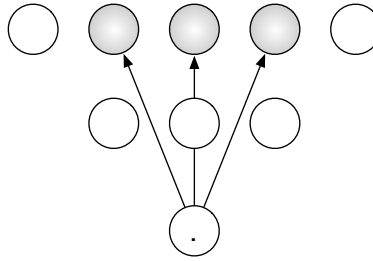


Fig. 7. The evaluation of the friend-of-a-friend expression yields a path from the root vertex to the vertex’s friends’ friends. This path can be interpreted as a virtual/inferred/implicit/derived edge. For the sake of diagram clarity, no edges are labeled. Assume that all edges are labeled “friend-of-a-friend.”

4 Conclusion

The concept of a graph was introduced in the late 19th century. During the many decades that followed, the world of graphs was primarily left to the toiling of mathematicians. In the last few decades, the sociology, physics, and computer science communities introduced a suite of algorithms and insightful realizations about the nature of graphs found in the real-world. Moreover, the increasingly large volume of data made available by the Internet has yielded datasets that reflect the graphs found in our technological and social systems. To satiate the need to handle and process these large-scale graphs, graph databases have come to the forefront. To make use of the graphs beyond simply representing their explicit structure, graph traversal frameworks and algorithms have been developed in order to shape graphs by driving the evolution of the entities that they model—e.g. humans and their relationships to one another and the objects of their world [11].

References

1. Franz Baader, Diego Calvanese, Deborah L. McGuinness, Daniele Nardi, and Peter F. Patel-Schneider, editors. *The Description Logic Handbook: Theory, Implementation and Applications*. Cambridge University Press, January 2003.
2. Ulrick Brandes and Thomas Erlebach, editors. *Network Analysis: Methodological Foundations*. Springer, Berlin, DE, 2005.
3. Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Hsieh, Deborah A. Wal-lach, Mike Burrows, Tushar Chandra, Andrew Fikes, and Robert E. Gruber. Bigtable: A distributed storage system for structured data. In *Proceedings of the 7th USENIX Symposium on Operating Systems Design and Implementation*, pages 205–218, Berkeley, CA, 2006. USENIX Association.
4. Edgar F. Codd. A relational model of data for large shared data banks. *Com-munications of the ACM*, 13(6):377–387, 1970.

5. Giuseppe DeCandia, Deniz Hastorun, Madan Jampani, Gunavardhan Kakulapati, Avinash Lakshman, Alex Pilchin, Swaminathan Sivasubramanian, Peter Voss hall, and Werner Vogels. Dynamo: Amazon's highly available key-value store. *SIGOPS Operating Systems Review*, 41(6):205–220, 2007.
6. Giorgio Gallo, Giustino Longo, Stefano Pallottino, and Sang Nguyen. Directed hypergraphs and applications. *Discrete Applied Mathematics*, 42(2-3):177–201, 1993.
7. Eric Miller. An introduction to the Resource Description Framework. *Bulletin of the American Society for Information Science and Technology*, 25(1):15–19, November 1998.
8. Eric Prud'hommeaux and Andy Seaborne. SPARQL query language for RDF. Technical report, World Wide Web Consortium, October 2004.
9. Marko A. Rodriguez and Peter Neubauer. The graph traversal pattern. [book chapter in review] arXiv:1004.1001, AT&Ti and NeoTechnology, 2010.
10. Marko A. Rodriguez and Joshua Shnavier. Exposing multi-relational networks to single-relational network analysis algorithms. *Journal of Informetrics*, 4(1):29–41, 2009.
11. Marko A. Rodriguez and Jennifer H. Watkins. Faith in the algorithm, part 2: Computational eudaemonics. In Juan D. Velásquez, Robert J. Howlett, and Lakhmi C. Jain, editors, *Proceedings of the International Conference on Knowledge-Based and Intelligent Information & Engineering Systems*, volume 5712 of *Lecture Notes in Artificial Intelligence*, pages 813–820. Springer-Verlag, 2009.
12. John F. Sowa. *Principles of Semantic Networks: Explorations in the Representation of Knowledge*. Morgan Kaufmann, San Mateo, CA, 1991.
13. Richard J. Trudeau. *Dots and Lines*. Kent State University Press, 1976.