Euclid and His Twentieth Century Rivals: Diagrams in the Logic of Euclidean Geometry

Nathaniel Miller

January 12, 2007

**CENTER FOR THE STUDY
OF LANGUAGE
AND INFORMATION**

"We do not listen with the best regard to the verses of a man who is only a poet, nor to his problems if he is only an algebraist; but if a man is at once acquainted with the geometric foundation of things and with their festal splendor, his poetry is exact and his arithmetic musical."

- Ralph Waldo Emerson, *Society and Solitude* (1876)

Contents

1

Background

In 1879, the English mathematician Charles Dodgson, better known to the world under his pen name of Lewis Carroll, published a little book entitled *Euclid and His Modern Rivals*. Dodgson was concerned by the fact that quite a number of different nineteenth century authors had written their own treatments of planar geometry, most claiming to improve on Euclid, and each one slightly different in the order of its theorems, in which theorems it chose to include, in the proofs given of these theorems, in its treatment of the theory of parallel lines, and in other aspects. Dodgson's book was written "[i]n furtherance of the great cause which I have at heart—the vindication of Euclid's masterpiece. . . ."[1] It is written mostly in the form of a dream dialogue between a nineteenth century mathematician, Minos, and the ghost of Euclid. In it, they consider each of the modern rivals in turn, and conclude in each case that, while many of the rivals have interesting things to say, none of them are a more appropriate basis for the study of a beginning geometry student than Euclid's *Elements*.

At the time at which Dodgson wrote his book, the subjects of geometry and logic were both entering a period of rapid change after having remained relatively constant for two thousand years. There had been enough change already to make Dodgson feel that Euclid needed defending. In the hundred and twenty-five years since then, however, there have been much larger changes in these fields, and, as a result, rather than just undergoing some small changes, Euclidean geometry in general, and Euclid's proofs in particular, have mostly fallen out of the standard mathematics curriculum. This is at least in part because Euclid's *Elements*, which was viewed for most of its existence as being the gold standard of careful reasoning and mathematical rigor, has come to be viewed as being inherently and unsalvageably informal and

[1] Dodgson (1885)

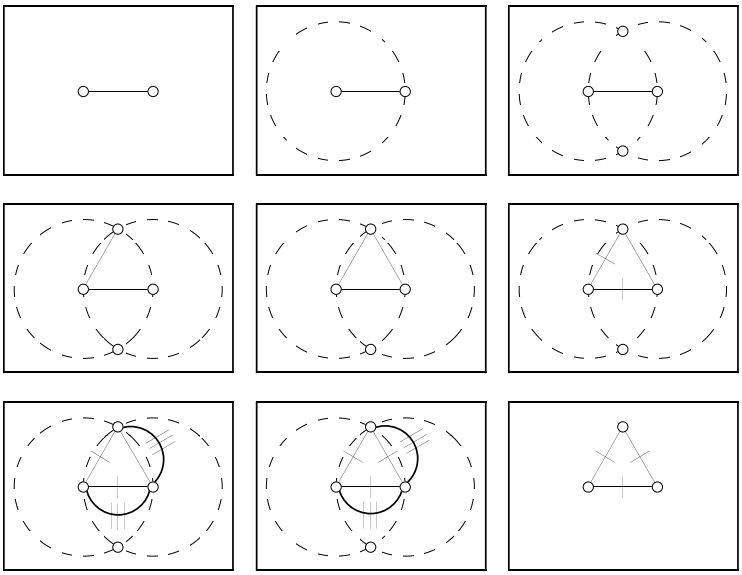


FIGURE 1 Euclid's first proposition.

unrigorous.

One key reason for this view is the fact that Euclid's proofs make strong use of geometric diagrams. For example, consider Euclid's first proposition, which says that an equilateral triangle can be constructed on any given base. While Euclid wrote his proof in Greek with a single diagram, the proof that he gave is essentially diagrammatic, and is shown in Figure 1.

In the decades following the publication of *Euclid and His Modern Rivals*, the field of mathematical logic blossomed, and a central idea in this blossoming was the movement towards increasing rigor. It was at this time that the idea of a formal proof was developed. A formal proof is one in which all of the rules that can be used are set out in advance so carefully as to leave no room for interpretation or subjectivity, and in which each step of the proof uses one of these rules. (We now have a more concrete way to describe such a system than was available when formal systems were first invented: a formal system is one whose rules are so concrete that they can be implemented by a computer.) This idea of a formal proof is a descendant of the idea which found its first enduring expression in Euclid's *Elements*: that proofs should proceed in logical sequence from axioms set out in advance. However, a view that came to be commonly accepted at this time and a view that is still

common is that, because of their use of diagrams, Euclid's proofs are inherently informal. Under this view, while diagrams may make proofs easier for students and others to follow, proofs that use diagrams cannot be formal because it isn't clear exactly what rules govern their use. The comments made by Henry Forder in *The Foundations of Euclidean Geometry* in 1927 are typical of this view: "Theoretically, figures are unnecessary; actually they are needed as a prop to human infirmity. Their sole function is to help the reader to follow the reasoning; in the reasoning itself they must play no part."[2] Most formal proof systems have therefore been sentential—that is, they are made up of a sequence of sentences in some formal language. It is easy to understand why: it is relatively easy to write down concrete rules that manipulate strings of symbols, in much the same way that algebraic equations are manipulated in high school algebra. Such a sentential axiomatization of geometry was given by David Hilbert in 1899, and since then, his axiomatization has replaced Euclid as the commonly accepted foundation of geometry.

However, while Hilbert's axiomatization has replaced Euclid's *Elements* as the theoretical basis for geometry, most informal geometric proofs still use diagrams and more or less follow Euclid's proof methods. The kinds of diagrams found in Figure 1 should be familiar to anyone who has ever studied planar geometry, and they follow standard conventions: points, lines, and circles in the Euclidean plane are represented by drawings of dots and different kinds of line segments, which do not have to really be straight, and line segments and angles can be marked with different numbers of slash marks to indicate that they are congruent to one another. The formal sentential proofs given in a system like Hilbert's are very different from these kinds of informal diagrammatic proofs. So, while Hilbert's system provides a formalization of the theorems of geometry, it doesn't provide a formalization of the use of diagrams or of many commonly used proof methods. It has been generally assumed that this is because these methods are inherently informal; but although they *have* not been previously formalized, this doesn't show that they *can* not be formalized. A natural question, then, is whether or not diagrammatic proofs like those in Euclid and like the one in Figure 1 can be formalized in a way that preserves their inherently diagrammatic nature.

The central aim of the present book is to show that they can. In fact, the derivation contained in Figure 1 is itself a formal derivation in a formal system called **FG**, which will be defined in the following sections

[2] (Forder, 1927, p.42)

of this book, and which has also been implemented in the computer system **CDEG** (Computerized Diagrammatic Euclidean Geometry). These systems are based a precisely defined syntax and semantics of Euclidean diagrams. We are going to define a diagram to be a particular type of geometric object satisfying certain conditions; this is the syntax of our system. We will also give a formal definition of which arrangements of lines, points, and circles in the plane are represented by a given diagram; this is the semantics. Finally, we will give precise rules for manipulating the diagrams—rules of construction, transformation, and inference.

In order to work with our diagrams, we will have to decide which of their features are meaningful, and which are not. A crucial idea will be that all of the meaningful information given by a diagram is contained in its topology, in the general arrangement of its points and lines in the plane. Another way of saying this is that if one diagram can be transformed into another by stretching, then the two diagrams are essentially the same. This is typical of diagrammatic reasoning, and, although it has not been previously treated formally, this idea has a long informal tradition in geometry. Proclus, the fifth century commentator on Euclid, writes that each case in a geometric proof "announces different ways of construction and alteration of positions due to the transposition of points or lines or planes or solids." (This is Sir Thomas Heath's translation, as given in Euclid (1956).) Thus, we see that the idea of considering cases to be different when the arrangements of the geometric objects being considered are topologically different is an ancient one.[3]

Our formalization of Euclid's proof methods is useful for several reasons: it lets us better understand his proofs; it allows us to prove metamathematical results about these kinds of proofs; and it shows that there is no inherent reason that the modern foundations of geometry must look completely different from the ancient foundations found in the *Elements*. Thus the aims of this book are not far removed from Dodgson's aims in 1879: to show that, while modern developments in logic and geometry may require changes in Euclid's development, his basic ideas are neither outdated nor obsolete.

[3]See Manders (1995) for an extended discussion of how ancient Greek proof practices made use of the topology of a diagram.

1.1 A Short History of Diagrams, Logic, and Geometry

In order to make sense out of this work, we need to put it in historical perspective.

The use of diagrams in geometry has a long history.[4] In fact, geometric diagrams are found among some of the oldest preserved examples of written mathematics, such as the Babylonian clay tablets found by archeological digs of ancient Mesopotamian city mounds at the end of the nineteenth century. These tablets, most of which are believed to date from around 1700 B.C., contain some fairly sophisticated arithmetical computations, and a number of them include diagrams. For example, the old-Babylonian tablet shown in Figure 2 (reproduced from Aaboe (1964)) shows the computation of the length of the diagonal of a square with sides of length 30, using a very good approximation of the square root of two. Geometric diagrams are also found in ancient Egyptian, Chinese, and Indian mathematical works.

It is with the Greeks, though, that mathematics really came into its own, and first and foremost among the Greek mathematical texts that have come down to us is Euclid's *Elements*. In fact, Euclid's *Elements* was such a seminal work that it has almost entirely eclipsed older Greek mathematical works—even though it wasn't written until around 300 B.C., long after the crowning achievements of the Greeks in art and literature, and thirty years after Alexander The Great had incorporated Greece into his empire centered in Alexandria, in Egypt. (In fact, Euclid himself lived and worked in Alexandria.) Thus, despite the fact that Euclid's *Elements* was part of a rich Greek mathematical tradition dating back to the beginning of the sixth century B.C., almost no earlier Greek mathematical works have come down to us in their entirety. This seems to be largely because *The Elements* suc-

[4]The history given here is not meant to be exhaustive, and is drawn from many sources. For more information, see Aaboe (1964) and Neugebauer and Sachs (1945) for discussion of Babylonian mathematics; Joseph (1991) for discussion of the mathematics of other ancient cultures and the development of algebra in Arabia; Kline (1967) for discussion of the history of Greece and Greek mathematics and a thumbnail sketch of the history of mathematics up to the twentieth century; Simmons (1985) and Bell (1937) for biographies of many mathematicians, including Archimedes, Descartes, Fermat, Leibniz, Gauss, Lobachevsky, and Boole; Bourbaki (1994), Kline (1980), and Dipert (2001) for the history of the discovery of non-Euclidean geometry, the arithmetization of mathematics, and the formalization of logic; Gardner (1982) and Shin (1994) for the history of logic diagrams; and Barwise and Allwein (1996) for the history of recent developments in the theory of reasoning with diagrams. Also see Greaves (2002), which is an excellent account of a large part of this history from a philosophical perspective.

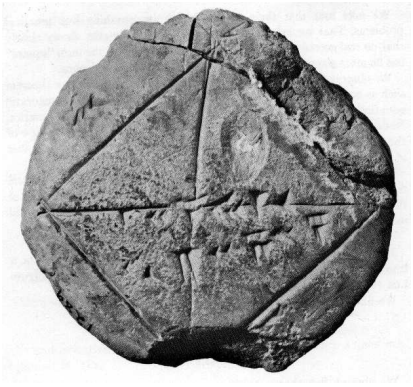


FIGURE 2 A Babylonian Tablet dating from around 1700 B.C.

ceeded in incorporating the majority of the preexisting mathematics into its logical development. *The Elements* has been a preeminent work in mathematics since the time it was written for a number of reasons, and among them is the fact that Euclid set down his assumptions in advance and tried to give explanations for why geometrical facts were true on the basis of his assumptions and previously shown facts. Thus, it is with Greek mathematics that we first encounter the notions of mathematical proof and the logical development of a subject. We also find a precursor of formal symbolic logic in the Greek theory of syllogistic reasoning, codified in Aristotle's *Prior Analytics*, written about fifty years before Euclid's *Elements*. Euclid's main concern in *The Elements* was Euclidean geometry, and, as we have already seen, his proofs of geometric facts rely heavily on diagrams. In fact, his first three postulates specify diagrammatic actions that can be performed in the course of a proof, although they are often translated in ways that obscure this fact: for example, his first postulate allows you "To draw a straight line from any point to any point." (See Appendix A for Sir Thomas Heath's literal translations of Euclid's Postulates.) Thus, these postulates, as originally stated, are hard to understand in any way that isn't essentially diagrammatic. This will be discussed at greater length later on.

The rise of the Roman Empire around 200 B.C. more or less eclipsed Greek culture, and in particular it eclipsed the Greek mathematical culture with its emphasis on proof. According to a famous story related by Plutarch,[5] the Greek mathematician Archimedes was killed during

[5]Plutarch (75)

the Roman conquest of the Greek city of Syracuse when he refused to come with an invading soldier until he was done studying a geometric diagram drawn in the sand. The British logician Alfred North Whitehead, one of the authors of the *Principia Mathematica*, thus remarked on the difference between the Greek and Roman cultures, "No Roman ever lost his life because he was absorbed in contemplation of a mathematical diagram."[6] As a result, we find fewer new developments in geometry or in logic for quite a long time after 200 B.C. Still, the *Elements* were always studied and carefully preserved, first by the Greeks and Romans, and then, after the destruction of Alexandria in 640 A.D., by Arabs in Arabic translations. The most important Arabic contribution to mathematics was probably their development of the subject of algebra. In fact, the word *algebra* comes from the arabic title of a book on the subject written by the Arabic mathematician Muhammad ibn Musa al-Khwarizmi in the ninth century A.D. In this work, al-Khwarizmi gives numerical methods for solving several different types of equations, followed by geometric proofs that these methods work. Thus, Arabic mathematics combined the subjects of algebra and geometry, using the Greek theory of geometry as the foundation for their developing theory of algebra.

It is not until the European Renaissance that we find steps away from the use of geometry as the foundation of mathematics. The first step came with the invention of analytic geometry in the 1630s by Pierre de Fermat and René Descartes. These men realized that it was possible to use algebra as a tool for studying geometry, and in doing so, they took the first steps towards a mathematics with arithmetic rather than geometry at its core. In Greek mathematics, geometry was viewed as the foundation for all other branches of mathematics, and so the Greek theories of arithmetic and algebra were based on their theory of geometry. The development of analytic geometry allowed mathematicians to instead base the theory of geometry on the theory of numbers, and thus it set mathematics on the path to arithmetization. The development of integral and differential calculus by Isaac Newton and Gottfried Leibniz independently in the 1660s and 1670s represented another big step in this direction, calculus being a tremendously powerful tool for studying geometric curves by using methods that are essentially arithmetical. The logical conclusion of this path was the definition of the real numbers in terms of the rationals, themselves defined in terms of the natural numbers, by Dedekind, Cantor, and others around 1870. With this development, geometry, with the real numbers at its core,

[6] Whitehead (1911)

could be seen as a mere extension of arithmetic. Thus, while Plato is quoted by Plutarch as having said that "God ever geometrizes,"[7] by the early 1800s this had become Jacobi's "God ever arithmetizes."[8]

Another factor that influenced the shift from geometry to arithmetic as the foundation of mathematics was the discovery of the consistency of non-Euclidean geometries in the 1820s by Gauss, Lobachevsky, and Bolyai. From the time of Euclid, students of *The Elements* had been unsatisfied with Euclid's fifth postulate. They felt that it was too inelegant and complex to be a postulate, and that it should therefore be possible to prove from the remaining postulates. Many proofs were proposed and even published, but each turned out to have made some additional assumptions. Finally, two thousand years after Euclid wrote, Gauss, Lobachevsky, and Bolyai each realized that there are consistent geometries in which the first four postulates hold, but the fifth does not. Thus, the fifth postulate cannot not be proven from the first four, because if it could, it would have to be true whenever they were. The discovery of these other geometries greatly weakened Euclidean geometry's claim to be the basis for all other mathematics. Before their discovery, it was thought that Euclidean geometry was just a codification of the laws of the natural world, and so it was a natural foundation on which to base the rest of mathematics. After people realized that other geometries were possible and that Euclidean geometry wasn't necessarily the true geometry of the physical world, it no longer had a claim to greater certainty than any other mathematical theory.

The transition from mathematics with geometry at its core to mathematics with arithmetic at its core had a profound influence on the way in which people viewed geometric diagrams. When geometric proofs were seen as the foundation of mathematics, the geometric diagrams used in those proofs had an important role to play. Once geometry had come to be seen as an extension of arithmetic, however, geometric diagrams could be viewed as merely being a way of trying to visualize underlying sets of real numbers. It was in this context that it became possible to view diagrams as being "theoretically unnecessary," mere "props to human infirmity."

As the rest of mathematics became arithmetized, so too did logic. The first steps in arithmetizing logic were taken Leibniz in the 1670s and 1680s, when he tried to develop a kind of algebraic system capturing Aristotle's rules for working with syllogisms. Leibniz's objective of finding a way of reducing syllogistic logic to algebra was finally realized

[7] Plutarch (1878)
[8] Bell (1937)

two hundred years later by George Boole in 1847. Over the next forty years various other people extended Boole's logical algebra in order to make it applicable to more of mathematics. Notable among them was the American Charles Sanders Peirce, who modified Boole's algebra to incorporate the use of relations and quantifiers. Finally, in 1879, the same year that *Euclid and His Modern Rivals* was published, Gottlob Frege published a book containing a logical system roughly equivalent to modern first-order predicate logic.

Interestingly, at the same time that these mathematicians were looking at ways to arithmetize logic, others were looking at ways to diagramize logic. The first method for using geometric diagrams of circles to solve syllogistic reasoning problems was given by Euler in 1761. His method of using circles to represent classes of objects was updated and improved by John Venn's introduction in 1880 of what are now known as Venn diagrams. Another diagrammatic system for representing logical statements was given in C. S. Peirce's system of Existential Graphs, introduced in 1897. (This is the same C. S. Peirce who had introduced quantifiers into Boole's algebra.) These Existential Graphs are notable not only for their expressive power, but also for the fact that Peirce gave a collection of explicit rules for manipulating them. Also worth mentioning here is Charles Dodgson, who in 1886 published a book called *The Game of Logic*, in which he proposed his own system of logic diagrams, equal in expressive power to those of Venn.

In the last decade of the nineteenth century, formal logic was well enough developed that careful axiomatizations of mathematical subjects could be given in formal languages. Around 1890, Giuseppe Peano published axiom systems for a number of mathematical subjects in a formal "universal" language that was based on the formalisms developed by Boole and Pierce. Among these were the axiomatization of arithmetic that now bears his name and an axiomatization of Euclidean geometry. Peano's axiomatization of geometry, along with several others, was eclipsed by David Hilbert's *Foundations of Geometry*, the first version of which was written in 1899.[9] By this point in time, Euclid's axiomatization and proofs had come to be seen as being insufficiently rigorous for a number of reasons, among them his use of diagrams. For example, the proof of Euclid's first proposition, shown in Figure 1 in the previous section, requires finding a point where the two circles intersect. Euclid seems to assume that this is always possible on the basis of the diagram, but none of his postulates appear to require the circles

[9]Hilbert's axioms from his *Foundations of Geometry* are reproduced in Appendix B.

to intersect. Hilbert's axiomatization was meant to make it possible to eliminate all such unstated assumptions. In fact, Hilbert showed that there is a unique geometry that satisfies his axioms, so that any fact that is true in that geometry is a logical consequence of his axioms. However, a proof from Hilbert's axioms may not look anything like Euclid's proof of the same fact. For example, Hilbert's axioms do not mention circles, so any proof of Euclid's first proposition will have to be very different from Euclid's proof.

Hilbert's axiomatization of geometry was part of a larger movement to try to put mathematics on the firmest possible foundation by developing all of mathematics carefully within formal systems that consisted of a small number of given axioms and rules of inference. This movement found its greatest expression in the *Principia Mathematica* of Bertrand Russell and Alfred North Whitehead, written between 1910 and 1913, which succeeded in developing a huge portion of mathematics from extremely simple axioms about set theory. However, it turned out that the goal of finding a finite set of axioms from which all of mathematics could be derived was impossible to achieve. In 1930, Kurt Gödel proved his First Incompleteness Theorem, which says approximately that no finite set of axioms is strong enough to prove all of the true facts about the natural numbers. The proof of this theorem involved translating logical statements into numbers and proofs into arithmetical operations on those numbers, and so it can be seen as having completed the arithmetization of logic. In any case, after Gödel's theorem was proven, logicians had to content themselves with more modest goals. In general, they still tried to reason from a small number of carefully specified axioms and rules of inference, because then if the axioms were true in a given domain and the rules of inference were sound, then any theorems proven would be correct.

It was not until recently that modern logic was applied to the study of reasoning that made use of diagrams. In the late 1980s, Jon Barwise and John Etchemendy developed a series of computer programs that were meant to help students visualize the concepts of formal logic. These programs, *Turing's World*, *Tarski's World*, and *Hyperproof*, included diagrams of a blocks world, and they inspired Barwise and Etchemendy to look more closely at forms of reasoning that used diagrams. In 1989, they published an article, "Visual Information and Valid Reasoning," reprinted in Barwise and Allwein (1996), which advanced the hypothesis that diagrammatic reasoning could be made as rigorous as traditional sentential reasoning and challenged logicians to look at diagrammatic reasoning more seriously.

Sun-Joo Shin, a student of theirs, began looking at the work that

had been done with logic diagrams a hundred years before. As we have seen, the development of systems of logic diagrams roughly mirrored the development of formal algebraic logical systems up to the end of the nineteenth century, but at that point they were for the most part abandoned as the theory of formal systems continued to develop in the twentieth century. Shin finally brought twentieth century developments in logic to bear on the theory of logic diagrams. She put together her own completely formal system of Venn diagrams, based on the earlier work of Euler, Venn, and Peirce, and showed that her system was both sound and complete—that the diagrams that could be derived from a given diagram system were exactly those that were its logical consequences. She also extended this system to include a more general form of disjunction and showed that the resulting diagrams had the same expressive power as the monadic first-order predicate calculus.

The first person to try to formalize the uses of diagrams in Euclidean geometry was Isabel Luengo, also a student of Jon Barwise. In her thesis,[10] finished in 1995, she introduced a formal system for manipulating geometric diagrams by means of formal construction and inference rules, and introduced the definition of "geometric consequence," which extends the notion of logical consequence to domains that include construction rules. However, her system does not incorporate the crucial idea that two diagrams should considered equivalent if and only if they are topologically equivalent, and as a result her system is unsound. For a detailed discussion of her formal system and an explanation of why it is unsound, see Appendix C.

In addition to the works just cited, in the time since Barwise and Etchemendy's article first appeared, many other people working in such varied fields as mathematics, philosophy, computer science, psychology, and cognitive science have taken up their challenge and have examined many different aspects of formal and informal reasoning with diagrams. Books written on aspects of diagrammatic reasoning include Barwise and Allwein (1996), Shin (1994), Hammer (1995), and Greaves (2002). Many interdisciplinary conferences on the subject have also been held, most with published proceedings; see, for example, Anderson et al. (2000), Hegarty et al. (2002), and Blackwell et al. (2004). Diagrammatic reasoning has thus become a growing field of inquiry.

1.2 The Philosophy Behind this Work

It should be stated from the outset that I am a mathematician, not a philosopher. However, the present work rests on and is motivated by

[10] Luengo (1995)

several philosophical stances that are worth articulating.

We are going to put together a formal system in which Euclid's proofs can be formalized. Why should we care if we can do this? As previously discussed, Euclid's *Elements* was seen as the gold standard in careful deductive reasoning from the time it was written until relatively recently, but it is now often viewed as being antiquated, inherently informal, and unsalvageable. So we might wonder which view is correct: are Euclid's methods of proof valid, or are they not? First off, let us note that it isn't quite clear what the question even means. In a formal system, we have a clearly defined notion of whether or not a proposed inference rule is valid—it is valid if it always gives a true conclusion when provided with true hypotheses. Euclid's proofs, however, are not part of a formal system in the modern sense, so we can't apply this test to them. This, by itself, doesn't mean that his methods are incorrect—only that they are informal. There are certainly informal proofs that mathematicians accept as being correct; in fact, practicing mathematicians almost never give proofs of their results in a formal system. Rather, they accept the idea that formal systems for doing mathematics exist, and if pressed might claim that their proofs could be translated into such a formal system. (This was the expressed purpose of the *Principia Mathematica* mentioned in the previous section.) So one possible meaning for the question, "Did Euclid use valid informal methods of proof?" is "Is it possible to create a formal system such that Euclid's informal proofs can be translated into formal proofs in this system?"

In some sense we are giving a kind of definition of what it means to have a valid informal method of giving proofs. We will call this the ***formality hypothesis***:

Hypothesis 1 (Formality Hypothesis) *An informal proof method is sound if and only if it is possible to give a formal system with the property that informal proofs using the informal methods can always be translated into equivalent correct proofs in the formal system.*

Several important points about this hypothesis should be made:

- It gives us a basis for evaluating informal methods of proof.
- It is inherently subjective and non-mathematical, because the question of what constitutes an "equivalent" formal proof is necessarily subjective.
- For this reason, it isn't possible to prove or disprove this hypothesis, and thus it is more like a definition than like a conjecture.
- It differs from a more traditional formalist position in that the basic objects being considered are proofs rather than statements. That is,

> we aren't just saying that a statement is true if a formal version of that statement can be proven in a formal system. We are saying that a proof of the statement is correct if the formal version of the statement can be proven by a formal version of *that proof* in a formal system.

In general, this hypothesis allows us to analyze the properties of an informal proof system by looking at the properties of a corresponding formal system.

Thus, the goal of this book is to give such a formal system in which Euclid's proofs can be duplicated. We will take the view that Euclid's proof methods should be viewed as being correct as long as they can be duplicated in our proof system. And we will prove meta-mathematical results about our formal system to try to shed some light on the properties of informal Euclidean geometric proofs. On the other hand, the existence of this formal system, showing that Euclid's proof methods can be formalized, provides support for the formality hypothesis.

One additional point that needs to be clarified here is the question of what we are willing to consider to be a formal system for the purposes of evaluating our hypothesis. The simplest possible definition that we can adopt here is that a system is completely formal if it can be completely implemented on a computer. This may seem obvious, but it means that the objects manipulated by the formal system can be anything at all as long as they can be translated into finite objects that a computer can manipulate. In the case of Euclidean geometry, this means that our formal system can manipulate diagrams, as long as we have some way representing the diagrams as finite objects in a computer.

Figuring out how to do this is non-trivial, and this step separates Euclidean geometry from other contexts in which people have formalized types of diagrammatic reasoning. For example, a Venn diagram consists of a finite number of regions each of which can contain an x, an o, both, or neither, and in which some of the xs can be connected to each other. If we know what is in each region and how they are connected, then we have all of the information contained in the diagram. It is clear that we can capture all of this information in a finite array. We'll have to do something similar for geometric diagrams, but figuring out how will be much more complicated. This is one reason that a skeptic might not believe ahead of time that it was possible to formalize the use of diagrams in geometry, even though the use of diagrams has been formalized in several other domains.

TABLE 1 Some of Euclid's definitions from Book I of *The Elements*.

Definitions

1. A point is that which has no part.
2. A line is breadthless length.
10. When a straight line set up on a straight line makes the adjacent angles equal to one another, each of the equal angles is ***right***, and the straight line standing on the other is called a ***perpendicular*** to that on which it stands.
15. A ***circle*** is a plane figure contained by one line such that the straight lines falling upon it from one point among those lying within the figure are equal to one another;
16. And the point is called the ***center*** of the circle.

1.3 Euclid's *Elements*

We are going to put together a formal system to formalize the kinds of proofs found in Euclid's *Elements*. Until the last century or so, any well-educated person would have studied Euclid's *Elements*, but since that is no longer the case, this section provides an introduction to that work. All of the quotations from the *Elements* given here are taken from Sir Thomas Heath's translation.[11]

Euclid's *Elements* consists of twelve books. The first four books are about elementary planar geometry; this is the part of the *Elements* that we would like to be able to formalize. Books five through ten are largely about the theory of ratio and proportion and about what would now be considered number theory; and books eleven through thirteen are about three dimensional geometry, and work their way up to the constructions of the five platonic solids. We will only be interested here in the kinds of proofs found in the first four books, which are the same kinds of proofs that are usually given in informal treatments of planar geometry today.

Euclid starts by giving twenty-three definitions, along with five postulates and five common notions. These postulates and common notions, along with several of the definitions, are given in Tables 1, 2, and 3. They are also reproduced, for the reader's convenience, in Appendix A.

The definitions that Euclid gives are really of two types. Some of his definitions, such as that of a circle in Definition 15, give specific defin-

[11] Euclid (1956)

TABLE 2 Euclid's Postulates from *The Elements*.

Postulates

Let the following be postulated:

1. To draw a straight line from any point to any point.
2. To produce a finite straight line continuously in a straight line.
3. To describe a circle with any center and distance.
4. That all right angles are equal to one another.
5. That, if a straight line falling on two straight lines make the interior angles on the same side less than two right angles, the two straight lines, if produced indefinitely, meet on that side on which are the angles less than the two right angles.

TABLE 3 Euclid's Common Notions from *The Elements*.

Common Notions

1. Things which are equal to the same thing are also equal to one another.
2. If equals be added to equals, the wholes are equal.
3. If equals be subtracted from equals, the remainders are equal.
4. Things which coincide with one another are equal to one another.
5. The whole is greater than the part.

ing properties that the object defined must have. These definitions are concrete and are sometimes used in his proofs. Other definitions, like that of a point given in Definition 1, give more general descriptions of the objects being described. They describe an object, but not in terms of properties that could be concretely used in a proof. A reader who already knows what a point is will probably agree that it "has no part," but a reader who didn't probably wouldn't find find this definition very useful. For this reason, modern treatments generally prefer to leave terms like this undefined. In such a treatment, we consider that the axioms could be referring to any model in which they are true, even if the points, lines, *etc.* might look very different in such a model than in what we consider to be the standard Euclidean plane. The fact that Euclid does give definitions of these terms shows that this isn't his intent: he has a particular model in mind, and gives an informal description of each kind of basic object in this standard model.

Euclid's Postulates and Common Notions set down the assumptions that he will use in his proofs. The Common Notions set down assumptions that are supposed to be true in general, while the Postulates set down the assumptions that are supposed to be particular to geometry. Thus, for example, Common Notion 1 states the transitive law of equality, which is still commonly taken as an axiom in formal treatments of many different fields of mathematics.

However, while some of Euclid's Postulates and Common Notions look like modern axioms, some do not. Look, for example, at his first postulate: "To draw a straight line from any point to any point." In Thomas Heath's literal translation, it isn't even a grammatical sentence, although other translations into English often obscure this fact. This postulate, like the next two, does not need to be a sentence, for the simple reason that it is not asserting a fact. Rather, it is stating a rule of the proof system: an allowable operation in this system is to draw a straight line from any given point to any other given point. This is an inherently diagrammatic rule, and it again makes it clear that the system that Euclid is defining here works in a significantly different way from most modern Hilbert-style formal deductive systems. In these systems, a set of axioms is usually given in first-order predicate logic, and then the first-order predicate calculus can be used to derive all of the logical consequences of these axioms. Euclid's first three postulates, as written, cannot be understood in this way, because rather than stating a fact assumed to be true, they state operations that can be performed. Many other translations, uncomfortable with this fact, change these construction axioms so that they do state facts. For example, the first postulate is often translated to match Hilbert's Axiom

I, 1: "For every two points A, B there exists a line a that contains each of the points A, B." The difference is subtle but important. The statement of an allowed rule in our proof system has been changed into a statement of fact in the underlying model.

Postulates two and three are also construction postulates. Postulate two means that any finite straight line segment can be extended indefinitely in either direction, making an infinite straight line. Postulate three says a circle can be drawn about any center point through any other point. Again, these specify allowed diagrammatic operations, which correspond to the operations that can be carried out with a straight edge and compass.

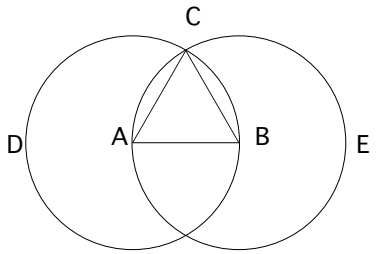
Postulate four seems trivial on a first reading, but really is not. Euclid's definition of a "right angle" is not one that is ninety degrees, but rather one that divides a straight line into two equal pieces. Thus, the postulate says that all such right angles, and therefore all straight lines, are the same everywhere. In effect, it says that the surface is uniform; it fails on surfaces like cones where some points are different than others.

The fifth postulate is the one that has been viewed by many readers through the ages as being too complicated to be taken as an axiom. The existence of non-Euclidean geometries in which it isn't true shows that it cannot be derived from the other postulates, however. We won't have a lot more to say about the fifth postulate here, except to note that, unlike most of the other proposed alternatives, Euclid's fifth postulate says nothing directly about parallel lines.

Now that we have looked at his definitions and axioms, we are in a position to look at one of Euclid's proofs. Euclid's proof of his first proposition is given in Table 4. The first thing to notice about this proposition is that, like the first three postulates, it gives a possible diagrammatic construction, rather than stating a fact to be shown. Just like his postulates, Euclid's propositions are of two types. Some state facts to be proven, while others, which are sometimes referred to as "Problems," give constructions which can be carried out using the construction postulates. Proposition 1 is one of these construction problems. Notice that Euclid first shows how to carry out the construction using the construction postulates, and then shows that the constructed triangle has the desired properties using other postulates, common notions, and definitions. Also notice that, although Euclid has written out his proof in words with a single diagram, the words really describe a sequence of diagrams. This is why, once we have constructed our formal system in which the sequence of diagrams shown in Figure 1 is a formal proof, it will seem reasonable to say that the sequence of

TABLE 4 Euclid's Proposition 1 from *The Elements*.

On a given finite straight line to construct an equilateral triangle.



Let AB be the given finite straight line.

Thus it is required to construct an equilateral triangle on the straight line AB.

With center A and distance AB let the circle BCD be described; [Post. 3]

again, with center B and distance BA let the circle ACE be described; [Post. 3]

and from the point C, in which the circles cut one another, to the points A, B let the straight lines CA, CB be joined. [Post. 1]

Now, since the point A is the center of the circle CDB, AC is equal to AB. [Def. 15]

Again, since the point B is the center of the circle CAE, BC is equal to BA. [Def. 15]

But CA was also proved equal to AB; therefore each of the straight lines CA, CB is equal to AB.

And things which are equal to the same thing are also equal to one another; [C.N. 1]

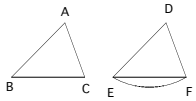
therefore CA is also equal to CB.

Therefore the three straight lines CA, AB, BC are equal to one another.

(Being) what it was required to do.

TABLE 5 Euclid's Proposition 4 from *The Elements*.

If two triangles have the two sides equal to two sides respectively, and have the angles contained by the equal straight lines equal, they will also have the base equal to the base, the triangle will be equal to the triangle, and the remaining angles will be equal to the remaining angles respectively, namely those which the equal sides subtend.



Let ABC, DEF be two triangles having the two sides AB, AC equal to the two sides DE, DF respectively, namely AB to DE and AC to DF, and the angle BAC equal to the angle EDF.

I say that the base BC is also equal to the base EF, the triangle ABC will be equal to the triangle DEF, and the angle ACB to the angle DFE.

For, if the triangle ABC be applied to the triangle DEF, and if the point A be placed on the point D and the straight line AB on DE, then the point B will also coincide with E, because AB is equal to DE.

Again, AB coinciding with DE, the straight line AC will also coincide with DF, because the angle BAC is equal to the angle EDF;

hence the point C will also coincide with the point F, because AC is again equal to DF.

But B also coincided with D; hence the base BC will coincide with the base EF.

[For if, when B coincides with E and C with F, the base BC does not coincide with the base EF, two straight lines will enclose a space: which is impossible.

Therefore the base BC will coincide with EF] and will be equal to it. [C.N. 4]

Thus the whole triangle ABC will coincide with the whole triangle DEF, and will be equal to it.

And the remaining angles will also coincide with the remaining angles and will be equal to them, the angle ABC to the angle DEF, and the angle ACB to the angle DFE.

Therefore *etc.*
(Being) what it was required to do.

diagrams is a formalization of the particular proof that Euclid gives, because the steps in the diagrammatic proof directly correspond to the steps that Euclid describes in his written proof.

One other proof that is worth looking at here is Euclid's proof of his Proposition 4, which is given in Table 5. This is his proof of the triangle congruence theorem usually referred to as Side-Angle-Side, or SAS for short. It says that if two triangles have an angle that is the same size in both, and such that the lengths of the sides that extend from this angle are the same size in both triangles, then the triangles must be congruent, meaning that one can be placed on the other so that they coincide, and that all of the other parts of the triangles must therefore be equal in size as well.

Note that part of the proof is enclosed in brackets; it is Heath's opinion that this piece is probably a later addition to the proof, and not in Euclid's original version.

Euclid proves this proposition by the method of superposition: he shows that if one triangle is placed on the other, then they will exactly coincide. Commentators on Euclid's *Elements* have widely objected to this method, as is discussed by Heath, for both philosophical and mathematical reasons. The mathematical reasons again mostly boil down to the idea that lots of assumptions have crept in here, and it is not clear precisely what the rules of using this method are. Euclid's stated assumptions say nothing about moving figures, although his common notion 4 is clearly included for the purpose of being cited here. Hilbert includes Side-Angle-Side as an axiom, rather than proving it. Many commentators also note that Euclid only rarely uses this method of proof, and, after having used it to prove Side-Angle-Side, tends to use Side-Angle-Side instead of using the method of superposition, and conclude that Euclid himself viewed the method as being suspect. Heath, for example, writes that "Euclid obviously used the method of superposition with reluctance. . . ."

We will show that this suspicion of the method of superposition is unfounded: our formal system will include rules that will allow us to use this method carefully, and, at the same time, will allow us to prove meta-mathematical results that will help explain why it is a method of proof to be used sparingly, even though it is valid.

2

Syntax and Semantics of Diagrams

As we have seen in the previous chapter, Euclid's proofs are inherently diagrammatic. Therefore, our first step in putting together a formal system to mimic Euclid's proofs will have to be to define precisely what we mean by a diagram in this context. Furthermore, as noted in Section 1.2, in order for our system to be completely formal and computerizable, we will need to give a definition with the property that all of the information contained in a diagram can also be encapsulated in a finite way in a computer.

2.1 Basic Syntax of Euclidean Diagrams

Figure 3 shows two examples of the sort of diagrams we want to consider. They contain dots and edges representing points, straight lines and circles in the plane, but note that a diagram may not look exactly like the configuration of lines and circles that it represents; in fact, it may represent an impossible configuration, like the second diagram in Figure 3.

Formally, we define a diagram as follows:

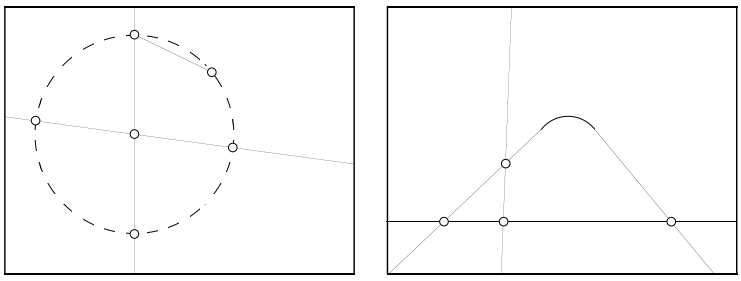


FIGURE 3 Two primitive diagrams.

Definition 1 A ***primitive Euclidean diagram*** D is a geometric object in the plane that consists of

1. a rectangular box drawn in the plane, called a ***frame***;
2. a finite set DOTS(D) of ***dots*** which lie inside the area enclosed by the frame, but cannot lie directly on the frame;
3. two finite sets SOLID(D) and DOTTED(D) of ***solid and dotted line segments*** which connect the dots to one another and/or the frame, and such that each line segment
   (a) lies entirely inside the frame,
   (b) is made up of a finite number of connected pieces that are either straight lines or else arcs of circles, which intersect each other only at their endpoints, and such that each of these pieces intersects at most one other piece at each of its endpoints,
   (c) does not intersect any other segment, any dot, the frame, or itself except at its endpoints, and
   (d) either forms a single closed loop, or else has two endpoints, each of which lies either on the frame or else on one of the dots;
4. a set SL(D) of subsets of SOLID(D), such that each segment in SOLID(D) lies in exactly one of the subsets; and
5. a set CIRC(D) of ordered pairs, such that the first element of the pair is an element of DOTS(D) and the second element of the pair is a subset of DOTTED(D), and such that each dotted segment in DOTTED(D) lies in exactly one of these subsets.

The intent here is that the primitive diagram represents a two-dimensional Euclidean plane containing points, straight lines and line segments, and circles. The dots represent points, the solid line segments in SOLID(D) represent straight line segments, and the dotted line segments in DOTTED(D) represent parts of circles. SL(D) tells us which solid line segments are supposed to represent parts of the same straight line, and CIRC(D) tells us which dotted line segments are supposed to represent parts of the same circle, and where the center of the circle is. (This comment is intended only to motivate the definitions being made now, and will be explained more carefully later on.) The sets in SL(D) are called ***diagrammatic lines***, or ***dlines*** for short, and the pairs in CIRC(D) are called ***diagrammatic circles*** or ***dcircles***. Elements of dlines are said to lie on the dline, and likewise, elements of the second component of a dcircle are said to lie on the dcircle; the first component of a dcircle is called the ***center*** of the dcircle. Each solid line segment must lie on exactly one dline, and each dotted line segment must lie on

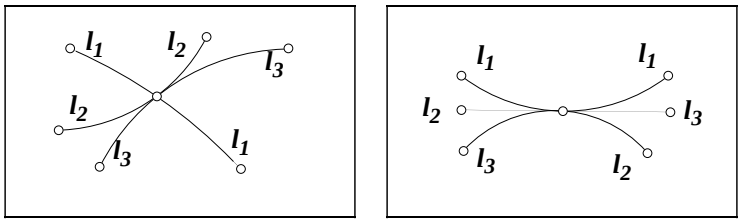


FIGURE 4 Examples of diagrammatic tangency.

exactly one dcircle. A dline or dcircle is said to intersect a given dot (or the frame) n times if it has n component segments with endpoints on that dot (or on the frame), counting a segment twice if both of its endpoints lie on the frame or on the same dot. Notice that it follows from the preceding definition that dlines and dcircles can only intersect other dlines and dcircles at dots (or on the frame, but this will eventually be disallowed).

We are now going to put some constraints on these diagrams to try to make sure that they look as much as possible like real configurations of points, lines, and circles in the plane. To begin, we would like to ensure that the dlines and dcircles come together at a dot in a way that mimics the way that real lines and circles could meet at a point. To this end, we first define the notion of diagrammatic tangency:

Definition 2 If each of e and f is a dcircle or dline that intersects the dot d exactly twice, then e and f are defined to be ***diagrammatically tangent*** (or ***dtangent***) at d if they do not cross each other at d.

This means that if s_{e1} and s_{e2} are the segments that are part of e which intersect d and, likewise, s_{f1} and s_{f2} are the segments from f that intersect d, then if s_{f1} occurs between s_{e1} and s_{e2} when the segments that intersect d are listed in clockwise order, then s_{f2} also occurs between s_{e1} and s_{e2} in this list. For example, in the first diagram in Figure 4, l_2 and l_3 are diagrammatically tangent to one another, while l_1 and l_2 are not. We are going to require the dcircles and dlines to intersect at d in such a way that the dtangency relation is transitive—in other words, so that if e and f intersect at d without crossing and f and g intersect at d without crossing, then e and g don't cross either (although they might both lie on the same side of f). This says that the situation in the second diagram in Figure 4, in which l_2 crosses l_3 but not l_1, cannot occur. Since dtangency is automatically symmetric and reflexive, this makes it into an equivalence relation. We can then extend the notion of diagrammatic tangency to dlines that only intersect d

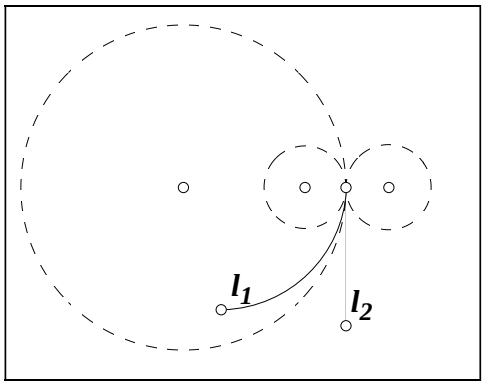


FIGURE 5 A non-viable primitive diagram

once by specifying that if e is such a dline, and e intersects d directly between two members of the same dtangency equivalence class, then e is dtangent to all of the members of that equivalence class. Thus, l_1 and l_2 in Figure 5 are dtangent to one another under this definition. A dline that only intersects d once is said to ***end*** at d.

We can now define a dot d to be viable as follows:

Definition 3 A dot is ***viable*** if

1. any dcircle that intersects the dot intersects it exactly twice;
2. any dline that intersects the dot intersects it at most twice;
3. the dcircles and dlines that intersect d do so in such a way so as to make the dtangency relation transitive; and
4. no two dlines are dtangent at d.

A primitive diagram D is viable if every dot in D is viable.

It follows from the preceding that if one member of a dtangency equivalence class crosses f at d, then all of the other members of the dtangency class also cross f at d; otherwise, some other member of the class would be dtangent to f at d, forcing them all to be dtangent to f at d. It also follows that each dtangency equivalence class can contain at most one dline, which may or may not end at d, since dlines are not allowed to be dtangent to other dlines. Notice that viability is a local property of diagrams—it says that the diagram is locally well-behaved at each dot. The two diagrams in Figure 3 are viable, while the three diagrams in Figures 4 and 5 are not. Note that our definition of viability allows viable diagrams to contain segments of lines, but not arcs of circles.

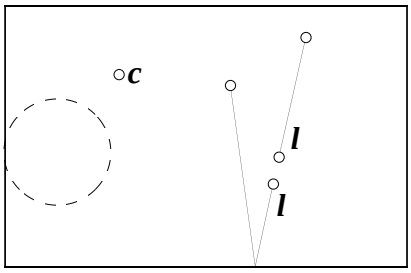


FIGURE 6 A viable diagram that isn't well-formed.

Next, we would like to ensure that the dlines and dcircles of our diagrams behave like real lines and circles. We do this with the following definition.

Definition 4 A primitive diagram D is ***well-formed*** if it is viable and

1. no dotted line segment in D intersects the frame;
2. no two line segments intersect the frame at the same point;
3. every dline and dcircle in D is connected—that is, given any two dots that a dline or dcircle P intersects, there is a path from one to the other along segments in P;
4. every dline has exactly two ends, where the ***ends*** of a dline are defined to be the points where it intersects the frame or a dot which it only intersects once; and
5. every dcircle in D is made up of segments that form a single closed loop such that the center of the dcircle lies inside that loop.

We call a dline that intersects the frame twice a ***proper dline***; one that intersects the frame once a ***d-ray***; and one that doesn't intersect the frame at all a ***dseg*** (not to be confused with the solid line segments that make it up). A well-formed primitive diagram is also called a ***wfpd***. Figure 6 shows a viable diagram that isn't well-formed and violates each of the four clauses of the definition. Both of the diagrams in Figure 3, however, are well-formed.

It should be noted that in principle, the diagrams drawn here should also tell you which segments make up each dline and dcircle. In the case of the first diagram in Figure 3, if we know that there are three dlines and one dcircle in this diagram, there are three different ways that the segments can be assigned to dlines and dcircles that make this a wfpd, as the reader should be able to check. Notice that if we are told that there are no dtangencies in a wfpd in which every dline is proper, then

there is only one way to assign segments to dlines and dcircles that is consistent with the diagram being well-formed, because you can determine which segments belong to the same dline or dcircle at a given dot by looking at the clockwise order in which the segments intersect the dot. In practice, it is usually clear which segments are intended to belong to the same dline or dcircle, and we won't indicate this unless it is unclear. We could also prove a theorem showing that every viable primitive diagram is equivalent to one in which two segments that intersect at a given dot are on the same dline iff they locally lie on a straight line, and are on the same dcircle iff they locally lie on some circle.

Finally, we have the following:

Definition 5 A primitive diagram is ***nicely well-formed*** if it is well-formed and

1. no two dlines intersect more than once;
2. no two dcircles intersect more than twice;
3. no dline intersects any dcircle more than twice;
4. if a dline is diagrammatically tangent to a dcircle, then they only intersect once;
5. if a dline intersects a dcircle twice, then the part of the dline that is between the two intersection points must lie on the inside of the dcircle; and
6. given any two non-intersecting proper dlines, if there is a third dline that intersects one of them, then it also intersects the other.

The last clause of this definition makes non-intersection of dlines an equivalence relation, and corresponds to the uniqueness of parallel lines. The first diagram in Figure 3 is nicely well-formed under two of the three assignments of segments to dlines and dcircles that make it a wfpd. The second diagram in Figure 3 is not nicely well-formed, since it contains two dlines that intersect twice. Nicely well-formed primitive diagrams are also called ***nwfpd***s.

Notice that the conditions for being viable are local conditions, the conditions for being well-formed are global conditions effecting individual dlines and dcircles, and the conditions for being nicely well-formed effect how dlines and dcircles can interact with one another globally.

2.2 Advanced Syntax of Diagrams: Corresponding Graph Structures and Diagram Equivalence Classes

We have now defined a primitive diagram to be a particular kind of geometric object. These diagrams contain somewhat too much information, though. We want the objects manipulated by our formal system to only contain a finite amount of information. As it stands, if we have a diagram containing only a single dot, that dot could be at any one of an infinite number of points inside the frame. We could specify its position by a pair of real numbers, but real numbers can have infinite decimal expansions, and therefore take an infinite amount of information to specify. We don't actually care exactly where the point is, however. We're not going to use the information about the specific location of the point. Instead, we're only going to use our diagrams to show the topology of how lines and circles might lie in the plane. For example, all of the diagrams that only contain a single dot have the same topology, even though each one is slightly different, so we'd like to consider them to all be the same diagram in some sense. In other words, we'd really like to look at equivalence classes of diagrams that contain the same topological information. In order to do this, we are going to define for each diagram an algebraic structure called a ***corresponding graph structure*** (abbreviated ***cgs***). The definition will be somewhat technical, but the idea is simple: the diagram's corresponding graph structure just abstracts the topological information contained in the diagram. Another way of saying this is that our definition will have the property that two diagrams will have isomorphic corresponding graph structures just if they have the same topological structure. As noted in Chapter 1, while this formal treatment is new, informal versions of this idea are ancient.

Some readers who are not interested in the technical details of this definition may wish to skip ahead to Definition 8, especially on a first reading.

A diagram D's corresponding graph structure will contain four kinds of information: a graph G that contains information about how the dots, frame, and segments intersect; for each point of intersection, information about the clockwise order in which the segments and frame intersect the point; for each doubly connected component DCC of G, a two-dimensional cell complex showing how the different regions of DCC (the connected components of the complement of DCC) lie with respect to one another; and for every connected component of G (except for the outermost component), information about which region of

the graph it lies in.

(Two vertices v_1 and v_2 in a graph G are said to be ***connected*** if there is a path from v_1 to v_2 in G, and they are said to be ***doubly connected*** if for any edge e of G, there is a path from v_1 to v_2 in the graph obtained from G by removing edge e. Being connected or doubly connected are equivalence relations, and their equivalence classes are called the connected or doubly connected components of G. The notion of a cell complex is a standard idea from algebraic topology; for reference, see any standard text book such as Hatcher (2002), which is freely available online.)

The notion of a cgs will be useful because we really want to think of two diagrams as being the same if they contain the same topological information, and so we will form equivalence classes of diagrams that have the same (isomorphic) corresponding graph structures. The corresponding graph structures are nice, constructive, algebraic objects that we can manipulate, reason formally about, or enter into a computer, rather than working directly with the equivalence classes. The data structures that **CDEG** uses to represent diagrams are essentially a version of these corresponding graph structures.

We start by defining the appropriate type of algebraic structure to capture the topology of a diagram.

Definition 6 A ***diagram graph structure*** S consists of

1. a set of vertices $V(S)$;
2. a set of edges $E(S)$;
3. for each vertex v in $V(S)$, a (cyclical) list $L(v)$ of edges from $E(S)$ (which lists in clockwise order the edges that are connected to v, telling us how to make the edges and vertices into a graph);
4. a two-dimensional cell-complex for each doubly connected component of the graph;
5. a function er_S from the non-outermost connected components of the graph to the two-cells of the cell-complexes (er stands for "enclosing region", and this function tells us which region each connected component lies in);
6. a subset DOTS(S) of $V(S)$;
7. two subsets of $E(S)$, called SOLID(S) and DOTTED(S);
8. a set SL(S) of subsets of $E(S)$; and
9. a set CIRC(S) of pairs whose first element is a vertex and whose second element is a set of edges.

We can now show how to construct a given diagram's corresponding graph structure. First note that the segments of a diagram D intersect

the frame in a finite number of points, which divide the frame into a finite number of pieces. We refer to these points as ***pseudo-dots*** and to these pieces as ***pseudo-segments***.

Definition 7 A diagram D's ***corresponding graph structure*** is a diagram graph structure S with the following properties:

1. $V(S)$ contains one vertex $G(d)$ for each dot or pseudo-dot d in D.
2. $E(S)$ contains one edge for every segment and pseudo-segment in D.
3. If d is any dot or pseudo-dot in D, then $L(G(d))$ lists the edges corresponding to the segments and pseudo-segments that intersect d, in the clockwise order in which the segments and pseudo-segments intersect d.
4. For each doubly connected component P of the graph G defined by $V(S)$, $E(S)$, and the lists $L(v)$, we define its ***corresponding cell complex*** C_P as follows:
   (a) C_P contains two-dimensional cells, one-dimensional cells, and zero-dimensional cells.
   (b) For each vertex v in P, C_P contains a corresponding 0-cell $C(v)$.
   (c) For each edge e of P, C_P contains a corresponding 1-cell $C(e)$.
   (d) Note that the segments and pseudo-segments of D that correspond to edges in P break up the plane into a finite number of connected regions, since there are only finitely many of them and they are piecewise arcs of circles and lines. Furthermore, because P is doubly connected, all but one of these (which we'll call the outer region) are simply connected. For each such simply connected region r, C_P contains a corresponding two-cell $C(r)$.
   (e) C_P is put together by connecting the zero-cells to the one-cells so that the boundary of $C(e)$ is the set containing $C(v_1)$ and $C(v_2)$ iff e connects v_1 and v_2 in G; and then attaching the two-cells to the resulting cell-complex so that the boundary of $C(r)$ is the loop that traverses $(C(G(s_1)), C(G(s_2)), \ldots, C(G(s_n)))$ in order if and only if the boundary of r in D consists precisely of $(s_1, s_2, \ldots, s_n)$ in clockwise order.
5. For each connected component p of G that does not contain the edges corresponding to the pieces of the frame, $er_S(p)$ is the unique two-cell $c = C(r)$ such that

   (a) the parts of D that correspond to p lie entirely in r, and
   (b) if they also lie entirely in a region r' corresponding to some other two-cell S, then r is contained in r'.
6. Each of the sets DOTTED(S), SOLID(S), DOTS(S), SL(S), and CIRC(S) is defined so that an element a of S is in one of these sets iff the corresponding element of D is in the corresponding set in D.

This definition now allows us to say what it means for two diagrams to contain the same information.

Definition 8 Two diagrams D and E are ***equivalent*** (in symbols, $D \equiv E$) if they have isomorphic corresponding graph structures.

This is an equivalence relation, and we normally won't distinguish between equivalent diagrams. If two diagrams D and E are equivalent, then there is a natural map f between the dots and segments of one diagram and the dots and segments of the other; we say that D and E are equivalent ***via*** f. If two graphs have corresponding graph structures that are isomorphic except that the orientations are all reversed, then we say that the diagrams are ***reverse equivalent***.

Next, we would like extend our notion of a geometric diagram to allow us to mark diagrammatic angles and segments as being congruent to other diagrammatic angles and segments. A ***diagrammatic angle*** or ***di-angle*** is defined to be an angle formed where two dlines intersect at a dot in a diagram. (They do not have to be adjacent to one another.) A ***marked diagram*** is a primitive diagram in which some of the dsegs and/or some of the di-angles have been ***marked***. A dseg is marked by drawing a heavy arc from one of its ends to the other and drawing some number of slash marks through it. If the dseg is made up of a single solid line segment, then it can also be marked by drawing some number of slash marks directly through the line segment. A di-angle is marked by drawing an arc across the di-angle from one dline to the other and drawing some number of slash marks through it. The arc and slash marks are called a marker; two dsegs or di-angles marked with the same number of slashes are said to be marked with the same marker. A single dseg or di-angle can be marked more than once by drawing multiple arcs.

We would also like our diagrams to be able express the existence of multiple possible situations. In order to show these, we will use diagram arrays. A ***diagram array*** is an array of (possibly marked) primitive diagrams, joined together along their frames. (It doesn't matter how they are joined.) Diagram arrays are allowed to be empty. Figure 7 shows a diagram array containing two different marked versions of the

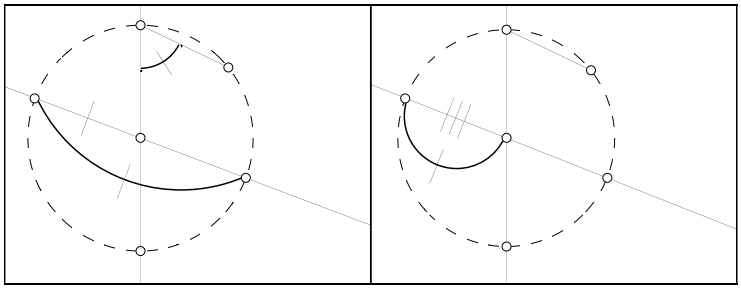


FIGURE 7 A diagram array containing two marked versions of the first primitive diagram in Figure 3.

first diagram in Figure 3.

We can extend our notion of diagram equivalence to marked diagrams and diagram arrays in the natural way. We define a ***marked diagram graph structure*** to be a diagram graph structure along with a new set MARKED whose elements are sets of dsegs and sets of ordered triples of the form <vertex, edge, edge>. We next define a marked primitive diagram D's corresponding marked graph structure to consist of the corresponding graph structure of D's underlying unmarked primitive diagram along with a set MARKED that for each segment marker in D contains the set of dsegs corresponding to the segments marked by that marker, and for each di-angle marker in D contains the set of triples $<v, e_1, e_2>$ such that the di-angle with vertex corresponding to v and edges corresponding to e_1 and e_2 in clockwise order is marked with that marker. Two marked diagrams are defined to be equivalent if and only if their corresponding marked graph structures are isomorphic; and two diagram arrays are equivalent if and only if there is a bijection f from the diagrams of one to the diagrams of the other that takes diagrams to equivalent diagrams.

2.3 Diagram Semantics

Now that we have carefully defined what a diagram is, we would like to discuss the relationship between diagrams and the real geometric figures that they represent. By a ***Euclidean plane***, we mean a plane along with a finite number of points, circles, rays, lines, and line segments designated in it, such that all the points of intersection of the designated circles, rays, *etc.* are included among the designated points. The elements of Euclidean planes are the objects that we would like to reason about. We consider the designated points of a Euclidean plane to divide its circles and lines into pieces, which we call ***designated***

edges.

It is very easy to turn a Euclidean plane P into a diagram. We can do this as follows: pick any new point n in P, pick a point p_l on each designated line l of P, and let m be the maximum distance from n to any designated point, any p_l, or to any point on a designated circle. m must be finite, since P only contains a finite number of designated points, lines and circles. Let R be a circle with center n and radius of length greater than m, and let F be a rectangle lying outside of R. Then if we let D be a diagram whose frame is F, whose segments are the parts of the edges of P that lie inside F, whose dots are the designated points of P, and whose dlines and dcircles are the connected components of the lines and circles of P, then D is a nwfpd that we call P's ***canonical (unmarked) diagram***. (Strictly speaking, we should say a canonical diagram, since the diagram we get depends on how we pick n and the p_l; but all the diagrams we can get are equivalent, so it doesn't really matter.) We can also find P's ***canonical marked diagram*** by marking equal those dsegs or di-angles in D that correspond to congruent segments or angles in P. These canonical diagrams give us a convenient way of saying which Euclidean planes are represented by a given diagram.

Definition 9 A Euclidean plane M is a ***model*** of the primitive diagram D (in symbols, $M \models D$, also read as "M satisfies D") if

1. M's canonical unmarked diagram is equivalent to D's underlying unmarked diagram, and
2. if two segments or di-angles are marked equal in D, then the corresponding segments or di-angles are marked equal in M's canonical marked diagram.

M is a model of a diagram array if it is a model of any of its component diagrams.

This definition just says that $M \models D$ if M and D have the same topology and any segments or angles that are marked congruent in D really are congruent in M. Note that this definition makes a diagram array into a kind of disjunction of its primitive diagrams and that the empty diagram array therefore has no models.

It is immediate from the definitions that every Euclidean plane is the model of some diagram, namely its canonical underlying diagram, and that if D and E are equivalent diagrams, then if $M \models D$, then $M \models E$. In other words, the satisfaction relation is well-defined on equivalence classes of diagrams. The full converse of this statement, that if $M \models D$ and $M \models E$, then $D \equiv E$, is not true, since D and

E may have different markings. However, it is true if D and E are unmarked. Also, if D is a primitive diagram that isn't nicely well-formed, then it has no models. To see this, notice that if $M \models D$, then D's underlying unmarked diagram D' is equivalent to M's canonical unmarked diagram, which is nicely-well formed; so D' is also nicely well-formed, as diagram equivalence preserves nice well-formedness, and so D is nicely well-formed since its underlying unmarked diagram is nicely well-formed.

We are going to use diagrams to reason about their models. In order to do this, we are going to define rules that will allow us to perform operations on given diagrams which return other diagrams. So we will need some way of identifying diagrammatic elements across diagrams. To do this, we can use a ***counterpart relation***, denoted $\mathrm{cp}(x, y)$, to tell us when two diagrammatic objects that occur in different primitive diagrams are supposed to represent the same thing. Formally, the counterpart relation is a binary relation that can hold between two dots or two sets of segments in any of the primitive diagrams that occur in some discussion or proof, but never holds between two dots or sets of segments that are in the same primitive diagram. Informally, people normally use labels to identify counterparts. For example, two dots in two different diagrams might both be labeled A to show that they represent the same point. The idea of a counterpart relation is due to Shin.[12]

[12]Shin (1994)

3

Diagrammatic Proofs

Now that we have a careful definition of what constitutes a diagram, we are ready to start putting together our formal system that will manipulate these diagrams.

3.1 Construction Rules

First of all, we need a way to use diagrams to model ruler and compass constructions like those found in the *Elements*. In order to do this, we will define several diagram construction rules. The rules work as follows: the result of applying a given rule to a given nwfpd D is a diagram array of (representatives of all the equivalence classes of) all the nwfpds that satisfy the rule (with corresponding parts of the diagrams identified by the counterpart relation). In effect, we are making sure that all of the different topological cases that could result when we apply the construction rule are considered. The new dlines and dcircles added by these rules are allowed to intersect any of the already existing dlines and dcircles, and the intersection points can be at new dots, as long as the resulting diagrams are still nicely well-formed. There will always be a finite number of resulting nwfpds, since each application of a rule will add a single new dot, dline, or dcircle, and the original diagram can only contain a finite number of dots and segments, none of which can be intersected more than twice by the new element, because of the conditions for niceness. We can apply the construction rules to diagram arrays by applying the rules to the individual primitive diagrams contained in the arrays. The diagram construction rules are given in Table 6.

Rule C3a is a special case of rule C3b, while C3b is derivable from C3a, as shown in Euclid's second proposition. Rules C1, C2, and C3a correspond to Euclid's first three postulates. Euclid's Postulates can be found in Appendix A.

TABLE 6 Diagram Construction Rules.

Diagram Construction Rules

C0. A dot may be added to the interior of any region, or along any existing segment, dividing it into two segments (unless the original segment is a closed loop, in which case it divides it into one segment).

C1. If there isn't already one existing, a dseg may be added whose endpoints are any two given existing distinct dots.

C2. Any dseg (or dray) can be extended to a proper dline.

C3a. Given two distinct dots c and d, a dcircle can be added with center c that intersects d if there isn't already one existing.

C3b. Given a dot c and a dseg S, a circle can be drawn about center c, with S designated to be a ***dradius*** of the dcircle. In general, we define a dseg to be a dradius of a dcircle if it is so designated by an application of this rule or if one of its ends lies on the dcircle and the other lies on the dcircle's center.

C4. Any dline or dcircle can be erased; any solid segment of a dline may be erased; and any dot that doesn't intersect more than one dline or dcircle and doesn't occur at the end of a dseg or dray can be erased. If a solid line segment is erased, any marking that marks a dseg or di-angle that it is a part of must also be erased.

C5. Any new diagram can be added to a given diagram array.

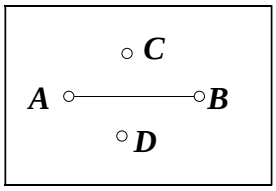


FIGURE 8 What can happen when points C and D are connected?

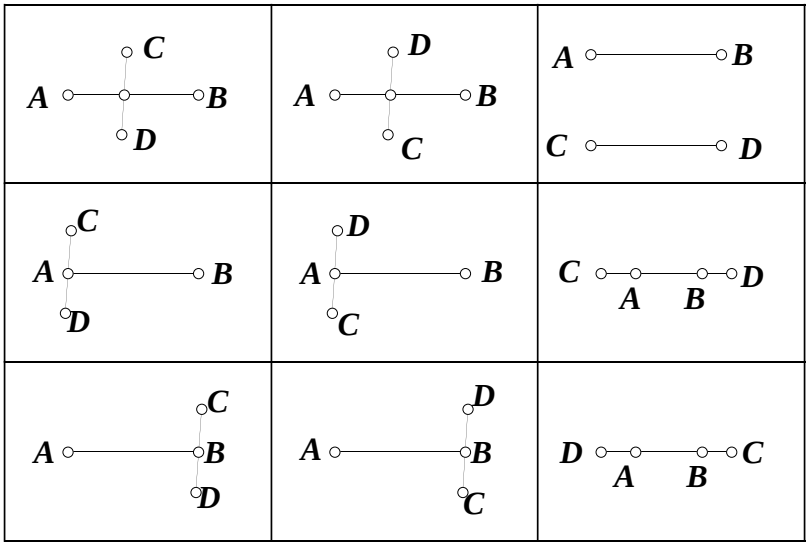


FIGURE 9 The result of applying rule C1 to points C and D in the diagram in Figure 8.

As a relatively simple example of how these rules work, consider the diagram shown in Figure 8. What happens if we apply rule C1 to this diagram in order to connect points C and D? We get the diagram array of all nwfpds extending the given diagram in which there is a dseg connecting points C and D. In this case, there are nine different topologically distinct possibilities, as **CDEG** confirms, which are shown in Figure 9. See Section 3.5 for a discussion of **CDEG** that includes a discussion of **CDEG**'s output in this case, with a detailed discussion of how it arrives at these nine possibilities.

A more useful example of these rules is given by the first four steps of the derivation of Euclid's first proposition shown in Figure 1, in which rule C3a is used twice, and then rule C1 is used twice. Notice that in this example, there is only one possible diagram that results from applying each of these rules. This is because many other possible diagrams have been eliminated because they are not nicely well-formed. For example, consider the step between the third and fourth diagrams in Figure 1. Call the points that are being connected A and C. The fourth diagram is supposed to be the array of all diagrams extending the third diagram in which A and C have been connected by a dseg (and nothing else has been added). It is, because there is only one such diagram, but if we had picked our rules for nice well-formedness

less carefully, there would have been others. Let's consider what would have happened if we had eliminated the fourth and fifth clauses in the definition of nice well-formedness (Definition 5), which say that if a dline intersects a dcircle twice, then the part of the dline that lies between the two intersection points must also lie inside the dcircle, and the dcircle cannot be dtangent to the dline at either of those points. Without these clauses, we would have gotten the array of ten diagrams shown in Figure 10. Thus, our definition of a nicely well-formed diagram saves us from considering many extra cases. Note that in this particular case, these extra diagrams could all be eliminated in one more step by using rule C2 to extend dseg AC into a proper dline. Since none of the extra cases can be extended in this way to give a nicely well-formed diagram (even without the fourth and fifth clauses of the definition), they would all have been eliminated.

A construction rule is said to be ***sound*** if it always models a possible real construction, meaning that if $M \models D$ and diagram E follows from D via this rule, then M can be extended to a model of E. The rules given in Table 6 are sound, because in any model, we can add new points, connect two points by a line, extend any line segment to a line, or draw a circle about a point with a given radius, and we can erase points, lines, and circles. In general, if every model M of D can be extended to a model of E, then we say that E is a ***geometric consequence*** of D, and write $D \models\!\subset E$. This definition of geometric consequence and the notation for it are due to Luengo.[13]

A diagram E is said to be ***constructible from diagram*** D if there is a sequence of diagrams beginning with D and ending with E such that each diagram in the sequence is the result of applying one of the construction rules to the preceding diagram; such a sequence is called a construction. Because our construction rules are sound, it follows by induction on the length of constructions that if E is constructible from D, then E is a geometric consequence of D.

The computer system **CDEG** uses explicit algorithms to compute the diagram graph structure that results from applying one of the construction rules to a given diagram, as is discussed in Section 3.5. These algorithms are based on the idea that if we want to know how a line can possibly continue from a given dot, it must either leave the dot along one of the already existing segments that leave the dot, or else it must enter one of the regions that the dot borders, in which case it must eventually leave that region at another dot or along another edge bordering the region, breaking the region into two pieces; along the

[13] Luengo (1996)

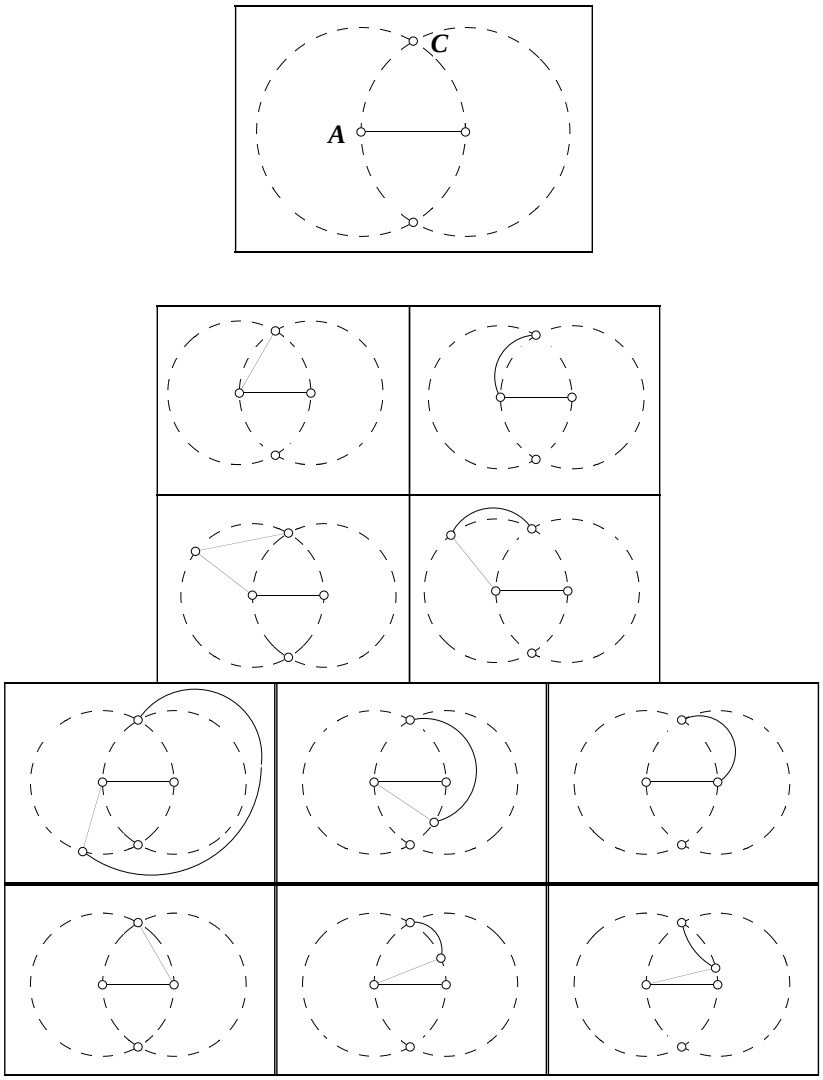


FIGURE 10 A modified construction.

way, it can intersect any of the pieces of any components that lie inside the region. This is reminiscent of Hilbert's axiom of plane order (II,4), which says that if a line enters a triangle along one edge, it must also leave the triangle, passing through one of the other two edges. In **FG**, this is a consequence of the definition of a nicely well-formed primitive diagram, rather than an explicitly stated axiom. This is typical: many of the facts that Hilbert adopts as his axioms of order and incidence are consequences of the diagrammatic machinery built into the definitions of **FG**.

In his commentary on Euclid's proof of Proposition 1,[14] Thomas Heath details several criticisms of Euclid's proof that have come down though history. They all have to do with additional assumptions that he would need to make in order to eliminate some of the possible extra unsatisfiable cases. For example, he writes, "It is a commonplace that Euclid has no right to assume, without premising some postulate, that the two circles *will* meet in a point C." We see though, that in our context, problems like these are taken care of by the underlying diagrammatic machinery, and therefore don't require a separate postulate. So it is possible to take the view that Euclid didn't, in fact, need any additional postulates here. He should, perhaps, have made it clearer what the rules governing his use of diagrams were. This is a subject, though, that he didn't address at all. It is in any case interesting to note that there is a fairly consistent use of diagrammatic machinery in Euclid, even if it is unremarked and possibly unconscious.

3.2 Inference Rules

Once we have constructed a diagram, we would like to be able to reason about it. For this purpose, we have rules of inference. Unlike the construction rules, when a rule of inference is applied to a single diagram, we get back a single diagram (at most). A rule of inference can be applied to a diagram array by applying it to one of the diagrams in the array. The rules of inference are given in Table 7. Rules R4 and R5 decrease the number of diagrams in a diagram array, and the other rules of inference leave that number constant, so applying rules of inference never increases the number of diagrams in a diagram array. If diagram (array) F can be obtained from E by applying a sequence of construction, transformation, and inference rules, then we say that F is ***provable*** from E, and write $E \vdash F$. (The transformation rules will be explained in the next section.)

Rules R1 and R2 correspond to Euclid's Common Notions 1 and 2,

[14] Euclid (1956)

TABLE 7 Rules of Inference.

Rules of inference

R1. (transitivity) If two dsegs or di-angles a and b are marked with the same marker and, in addition, a is also marked with another marker, then b can also be marked with the second marker.

R2. (addition) If there are four dsegs or di-angles a, b, c, and d such that a and b don't overlap and their union is also a dseg or di-angle e , and c and d don't overlap and their union is a dseg or di-angle f, then if a and c are marked with the same marker, and b and d are marked with the same marker, then e and f can be marked with the same new marker not already occurring in the given diagram.

R3. Any two dradii of a given dcircle may be marked with the same new marker.

R4. Given a diagram array that contains two diagrams that are copies of one another, one of them may be removed.

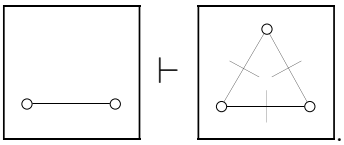
R5a. (CS) If a diagram contains two dsegs, one of which is properly contained in the other, and both of which are marked with the same marker, then it can be removed from a diagram array.

R5b. (CA) If a diagram contains two di-angles, one of which is properly contained in the other, and both of which are marked with the same marker, then it can be removed from a diagram array.

R6. Any dseg or di-angle can be marked with a single new marker. Any marker can be removed from any diagram.

to Hilbert's Axioms III, 2 and III, 3, and to Luengo's inference rules R4.5 and R4.4. Rules R5a and R5b correspond to Euclid's fifth common notion. Here, "CS" stands for "Contradiction of Segments" and "CA" stands for "Contradiction of Angles." Hilbert assumes R5b as his axiom III, 4, and uses it to prove R5a from SAS as we will show how to do in Section 4.3, while Luengo incorporates a version of R5a into her definition of syntactic contradiction. (See Luengo (1996), Hilbert (1971), and Appendices A and B.) R3 corresponds to Euclid's Definition 15. We have already incorporated a version of the uniqueness of parallel lines into our definition of nice well-formedness, but we could just as well have added it here. Euclid's fourth postulate is derivable from our other rules using the symmetry transformations, and Euclid's fifth postulate is derivable from the uniqueness of parallel lines in the usual way.

The second half of the proof in Figure 1 uses these inference rules. Beginning with the fifth diagram in the proof, we can apply rule R3 twice and rule R1 once to obtain a diagram in which all three sides of the triangle are marked equal, and then using R7 and C4 we can erase the extra markings and the circles, leaving just the triangle. Thus, Figure 1 shows that



.

Call the first diagram here A, and the second B. Since A is certainly constructible from the empty primitive diagram, B is also provable from the empty primitive diagram. (We write this as "$\vdash B$".) Notice that, unlike what we're used to with linguistic systems, $A \vdash B$ is actually a stronger statement than $\vdash B$, since diagrams A and B are related by the counterpart relation. So $\vdash B$ says that an equilateral triangle can be constructed, whereas $A \vdash B$ says that given any segment, an equilateral triangle can be constructed along that segment. Strictly speaking, $A \vdash B$ just means that we can get from A to B using our rules, and it is $A \mid\!\subset B$ that means that an equilateral triangle can be constructed along any given segment. But it is an immediate consequence of the soundness of our rules that if $A \vdash B$, then $A \mid\!\subset B$. It is easy to check that our rules are indeed sound. For example, to check that rule R1 is sound, assume that we are given two diagrams D and E such that $D \vdash E$ via rule R1. Then E differs from D only in that there are two dsegs or di-angles a and b in D and E such that in D, a is marked with two markings m and

n but b is only marked with marking m, while in E, b is also marked with marking n. Since E differs from D only in that b is marked with marking n in E, to show that $D \models E$ it suffices to show that if M is a model of D and o is any element of D that is marked with marking n, then the pieces of M that correspond to o and b are congruent. Since M is a model of D, the pieces of M that correspond to a and b are congruent, since they are both marked with marking m in D, and the pieces that correspond to a and o are congruent, since they are both marked by marking n in D; so the pieces that correspond to o and b are also congruent in M since congruence is a transitive relation in any Euclidean plane. So $M \models E$, which means that $D \models E$. The proofs that the other rules are sound are similar exercises in chasing definitions and then using a corresponding semantic fact about the models.

3.3 Transformation Rules

In order to formalize Euclid's method of superposition, we need to be able to use diagrams to model isometries: translations, rotations, and reflections. To do this, we first need the notion of a ***subdiagram***. A primitive diagram A is a subdiagram of B if A is constructible from B using only rule C4. Next, we define a diagram T to be an ***super transformation diagram*** of A in D (via transformation t) if A is a subdiagram of D, D is a subdiagram of T, and there exists another diagram B and a function $t : A \to B$ such that B is also a subdiagram of T, and A and B are equivalent or reverse equivalent diagrams via the map t. T is a ***transformation diagram*** of A in D via t if T is an super transformation diagram of A in D via t, and no proper subdiagram S of T is still a super transformation diagram of A in D via t. If A and B are equivalent, then it is an ***unreversed*** transformation diagram, and if they are reverse equivalent, then it is a ***reversed*** transformation diagram. Now we can incorporate symmetry transformations into our system by adding the rules in Table 8. Note that simple rotations and translations are special cases of rule S1, and reflections are a special case of rule S2.

Each of these rules, like the construction rules, always yields a finite number of consequences when applied to a single diagram. This is because the unmarked diagram array that results from applying one of these rules and then erasing all markings is a subarray of the the array that is obtained by constructing a copy of A in the appropriate spot in D using the construction rules.

The system that contains the construction rules C0–C4, the transformation rules S1 and S2, and the rules of inference R1–R6 is called

TABLE 8 Transformation Rules

Transformation Rules

S1. (glide) Given a diagram D, the subdiagram A, a dot a and a dseg l_1 ending at a in A, and a dot b and a dseg l_2 ending at b in D, the result of applying this rule is the diagram array of all unreversed transformation diagrams of A in D such that $t(a) = b$ and $t(l_1)$ lies along the same dline as l_2, on the same side of b as l_2.

S2. (reflected glide) Given a diagram D, the subdiagram A, a dot a and a dseg l_1 ending at a in A, and a dot b and a dseg l_2 ending at b in D, the result of applying this rule is the diagram array of all reversed transformation diagrams of A in D such that $t(a) = b$ and $t(l_1)$ lies along the same dline as l_2, on the same side of b as l_2.

FG (for "Formal Geometry").

As an example of how these transformation rules work, consider the diagram found in Figure 11. It is a logical consequence of this diagram that EF is congruent to BC, and we should therefore be able to mark it with three slash marks. This is one particular case of the rule of inference SAS, which we have already encountered as Euclid's Proposition 4:

SAS. If a diagram contains two triangles, such that two sides of one triangle and the included di-angle are marked the same as two sides and the included di-angle of the other triangle, then the remaining sides of the triangles can be marked with the same new marker, and each of the remaining di-angles of the first triangle can be marked the same as the corresponding di-angle of the other.

In **FG**, SAS is a derived rule; it can be derived from our symmetry transformations along with CA and CS. The proof is essentially identical to Euclid's proof, with a lot of tedious extra cases showing all of the ways that the triangles could possibly intersect. The idea is to move the two triangles together using the symmetry transformations and to then check that they must be completely superimposed.

In **FG**, the proof of this case of SAS has two steps. The first step is to apply rule S1 to the diagram in Figure 11, moving triangle ABC so that A' (= t(A)) coincides with D, and so that the image $A'B'$ of AB lies along DE. The possible cases that result are shown in the diagram arrays in Figures 12 and 13. For the sake of readability, many of the

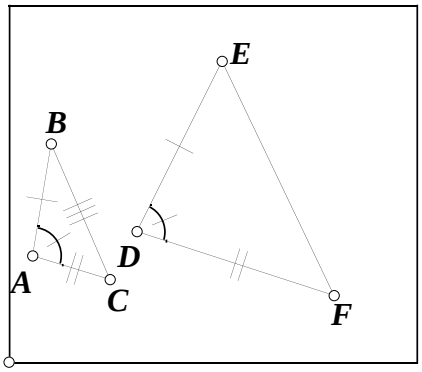


FIGURE 11 The hypothesis diagram for one case of SAS.

markings have been left off these diagrams, although all markings that are later needed have been left. Also, properly speaking, these figures are only some of the cases that are given by rule S1, because any part of $A'B'C'$ that lies outside of DEF can intersect ABC in any one of a number of ways. But the diagrams do show all of possible cases in which $A'B'C'$ doesn't intersect ABC.

The second step is to remove all of the extra cases using the rules of inference CA and CS. All of the diagrams shown in Figure 12 except for the very first one can be eliminated by applying CS to $A'B'$ and DE. In Figure 13, all of the diagrams in the first four rows and the first two diagrams in the fifth row are eliminated in the same way. The rest of the diagrams can be eliminated by using CA, except for the last two diagrams, which can also be eliminated by using CS. The cases that weren't shown in Figures 12 and 13, in which $A'B'C'$ intersects ABC, can also all be eliminated using CS and CA. Thus, we have shown SAS for one particular case, in which the two original triangles don't intersect and have the same orientation. The proof for the other cases is similar.

3.4 Dealing with Areas and Lengths of Circular Arcs

FG is the formal system that includes all the rules that have been defined in the preceding sections: the rules of construction, inference, and transformation. The system defined in this way is sufficient to do a lot of informal geometry, and strong enough to emulate most of the first three quarters of Book I of Euclid's *Elements*. However, after that point, Euclid starts giving proofs that involve comparing areas in diagrams. **FG**, as we have defined it so far, lacks a way to talk about equality of

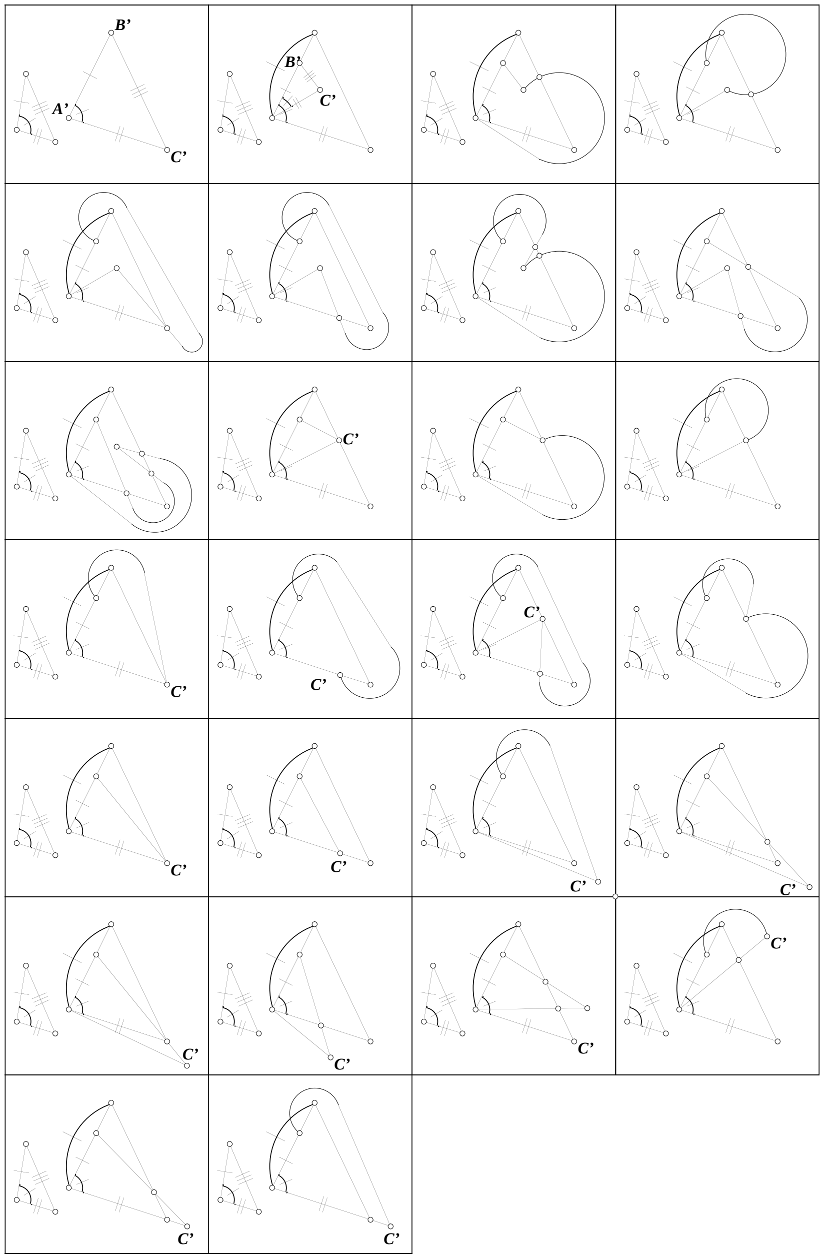


FIGURE 12 The first half of the cases that result from applying rule S1 to the diagram in Figure 11.

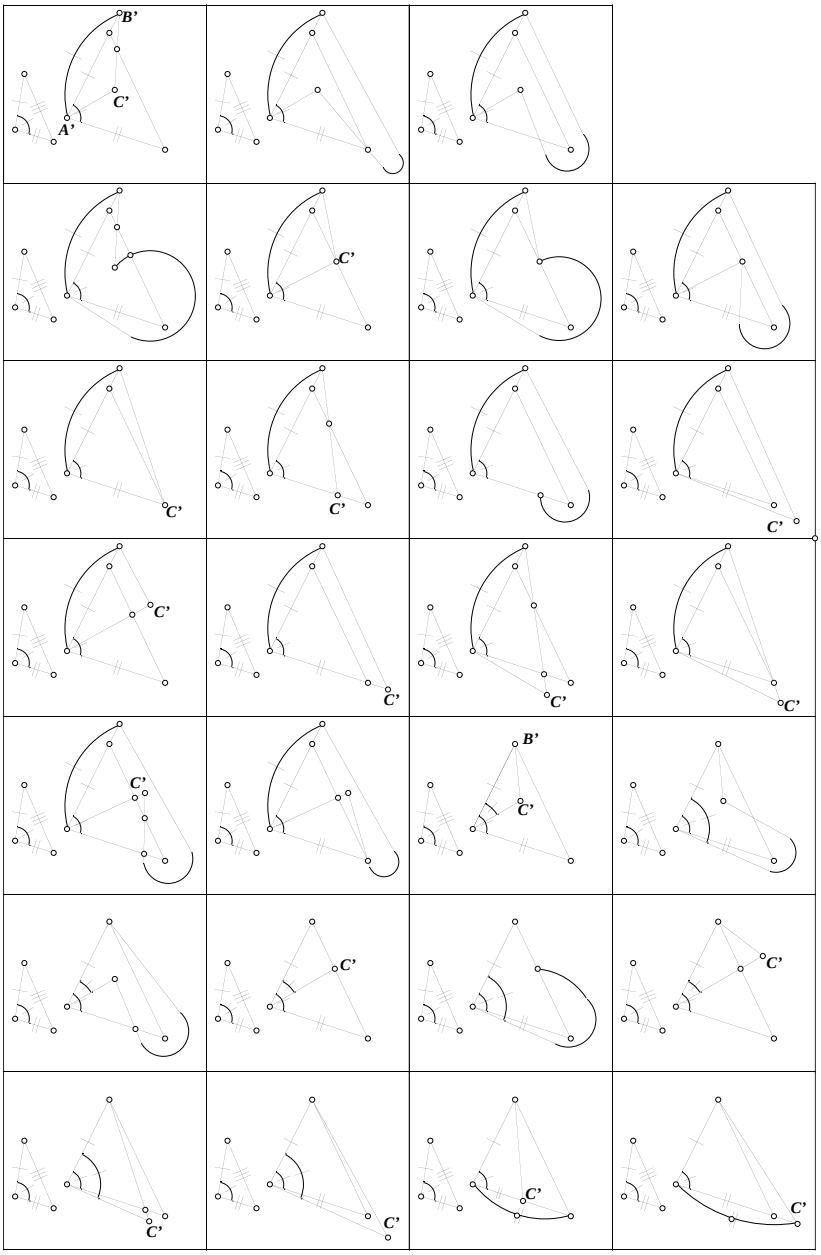


FIGURE 13 The second half of the cases that result from applying rule S1 to the diagram in Figure 11.

areas. Furthermore, in parts of Book III, Euclid compares the lengths of circular arcs. Again, **FG** lacks a way to compare lengths of arcs.

This difficulty is easily rectified. We can treat areas of regions and lengths of circular arcs in almost exactly the same way that we have treated lengths of segments and sizes of angles. We will now define such a system, which we will call **FG′**.

First of all, we can define a ***di-arc*** to be an arc on a dcircle, composed of one or more dotted segments; and we can define a ***dregion*** to be a set of connected regions in a diagram, none of which borders the frame.

Next, we need to allow these di-arcs and dregions to be marked congruent to other di-arcs and dregions, just as di-angles and dsegs can be. Formally, this is easy; we can just change the definition of the set MARKED that is part of a marked diagram graph structure to also include sets of di-arcs and dregions. Graphically, however, there isn't a standard informal way to mark regions as having the same area. Di-arcs can be marked with slash marks in the same way as dsegs, even though this isn't standard. The most common way to designate areas as being equal in informal treatments of geometry is to simply write this down in the text that accompanies the diagram. This is the approach that we will adopt: a dregion will be marked with a particular marker by noting this fact in text attached to a given diagram. Some people might argue that this makes our new system less than totally diagrammatic, but it is consistent with standard informal practice; formally, it looks just like our treatment of dsegs and di-angles; and we could certainly invent a purely diagrammatic way of representing this information, but it would probably make reading the diagrams much more complicated. This approach of including markings as separate text is actually the approach adopted in **CDEG** for all of its markings, even those of dsegs and di-angles. See Appendix D for more details.

In including text with our diagrams, we are moving closer to what is usually called ***heterogenous*** reasoning—reasoning that includes both diagrammatic and sentential information. Heterogenous reasoning has received a fair amount of attention in the literature on diagrammatic reasoning, particularly in discussions of Barwise and Etchemendy's Hyperproof system, which is a true heterogenous system. (See, for example, Barwise and Allwein (1996).) Our system isn't truly heterogenous, because the text that we're adding is just part of a diagram, rather than a separate sentence. Many informal treatments of geometry are arguably heterogenous, however, because, like Euclid's *Elements*, their arguments include both diagrams and written arguments. Certainly, it would be very interesting to try to extend **FG′** into a formal system that keeps **FG′**'s diagrammatic machinery, but also allows sentential

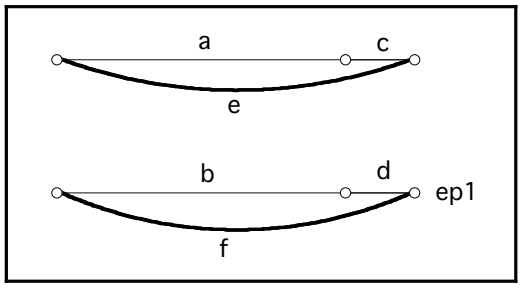


FIGURE 14 Proving the subtraction rule.

conclusions to be drawn from the diagrams. This is a possible direction for further work in this field.

Once we have the ability to mark di-arcs and dregions, we need to add rules of inference to use with them; these new rules are strictly analogous to the rules that we adopted for dsegs and di-angles. The new rules of inference are shown in Table 9. (Rules R3, R4, R5a, and R5b are the same in **FG′** as in **FG** and are therefore omitted from Table 9.) "CCA" stands for "Contradiction of Circular Arcs," and "CR" stands for "Contradiction of Regions."

The only really new rule here is rule R2.5 (subtraction), which corresponds to Euclid's Common Notion 3. We didn't need this rule for dsegs, di-angles, or di-arc, because in those cases the subtraction rule is derivable from the other rules. For example, assume that a, b, c, d, e, and f are dsegs such that a and c together make e, and b and d together make f, and we know that a is congruent to b and e is congruent to f. Furthermore, let the endpoint of d on the side away from b be called $ep1$. This configuration is shown in Figure 14. Then to show that c is congruent to d, construct a segment with the same length as c next to b on the same side as d, ending at dot $ep2$. (We can do this either by using circles, or else by using the transformation rules.) By the addition rule, we must have $b+c = a+c$. So if $ep1$ doesn't coincide with $ep2$, we will have a contradiction by rule CS. This means they must coincide, which makes c congruent to d, as required.

The reason that this argument doesn't work for areas is that, unlike segments and angles, two areas can fail to coincide when lined up without either being properly contained in the other. Another important difference to notice here is that segments and angles marked with the same marker must actually be congruent, in the sense that one can be placed exactly upon the other. This is not true of of regions.

TABLE 9 New and Modified Rules of Inference of **FG′**.

New and Modified Rules of Inference of FG′

R1. (transitivity)If two dsegs, di-angles, di-arcs, or dregions a and b are marked with the same marker and, in addition, a is also marked with another marker, then b can also be marked with the second marker.

R2. (addition) If there are four dsegs, di-angles, di-arcs, or dregions a, b, c, and d such that a and b don't overlap and their union is also an object of the same type e , and c and d don't overlap and their union is an object of the same type f, then if a and c are marked with the same marker, and b and d are marked with the same marker, then e and f can be marked with the same new marker not already occurring in the given diagram.

R2.5. (subtraction) If there are four dregions a, b, c, and d such that a and b don't overlap and their union is also a dregion e , and c and d don't overlap and their union is a dregion f, then if a and c are marked with the same marker, and e and f are marked with the same marker, then c and d can be marked with the same new marker not already occurring in the given diagram.

R5c. (CCA) If a diagram contains two di-arcs, one of which is properly contained in the other, and both of which are marked with the same marker, then it can be removed from a diagram array.

R5d. (CR) If a diagram contains two dregions, one of which is properly contained in the other, and both of which are marked with the same marker, then it can be removed from a diagram array.

R6. Any dseg, di-angle, di-arc, or dregion can be marked with a single new marker. Any marker can be removed from any diagram.

Two regions with the same area may have very different shapes. For the addition rule to be sound, we have to interpret the fact that two regions are marked with the same marker to only mean that they have the same area, not that they are actually congruent. Note that something similar is true for di-arcs: two circular arcs with the same length may none the less have different shapes. We will adopt the opposite convention for the di-arcs, however. If two di-arcs are marked with the same marker, we will require that they must be actually congruent. Both of these conventions are consistent with Euclid's use of the word "equals." For example, Euclid's Proposition 35 in Book I of the *Elements* is "Parallelograms which are on the same base and in the same parallels are equal to one another." These parallelograms are equal in area, but do not have the same shape. On the other hand, he never asserts the equality of circular arcs which are equal in length but not congruent, and in his proof of Book III, Proposition 23 ("On the same straight line there cannot be constructed two similar and unequal segments of circles on the same side"), he immediately takes "unequal" to mean "non-coinciding."

As a simple example of these new rules, consider the proof of the case of SAS shown in Figures 11, 12, and 13. If we add a new dregion marker to the region inside triangle ABC in Figure 11, then after applying the transformation rule and eliminating all of the extra cases, we will have the additional conclusion that the areas of the two triangles are equal, in addition to the sides and angles. As a more complicated example, consider the proof of one case of Euclid's Proposition 35 from Book I of the *Elements*, as shown in Figure 15. This proof assumes two additional rules. The first one is given by the previous proposition (I.34), which says that opposite sides of parallelograms are congruent. The other is the fact that when parallel lines are cut by a transversal, the corresponding angles are congruent, which is Proposition 29 of Book I.

In the derivation shown in Figure 15, some of the steps have been combined, and some of the segment length markers are reused. The fact that we are trying to prove is that the two parallelograms $ABCD$ and $EBCF$ have the same areas. The second diagram follows from the first by two applications of Euclid's proposition I.34, which says that opposite sides of parallelograms are congruent. The third diagram follows from the second by the transitivity rule R1. The fourth follows from the third by rule R6. The fifth follows from the fourth by the addition rule (R2) for dsegs. The sixth follows from the fifth by Euclid's Proposition I.29 about corresponding angles of parallel lines, and the seventh follows from the sixth by another application of I.34. The eighth

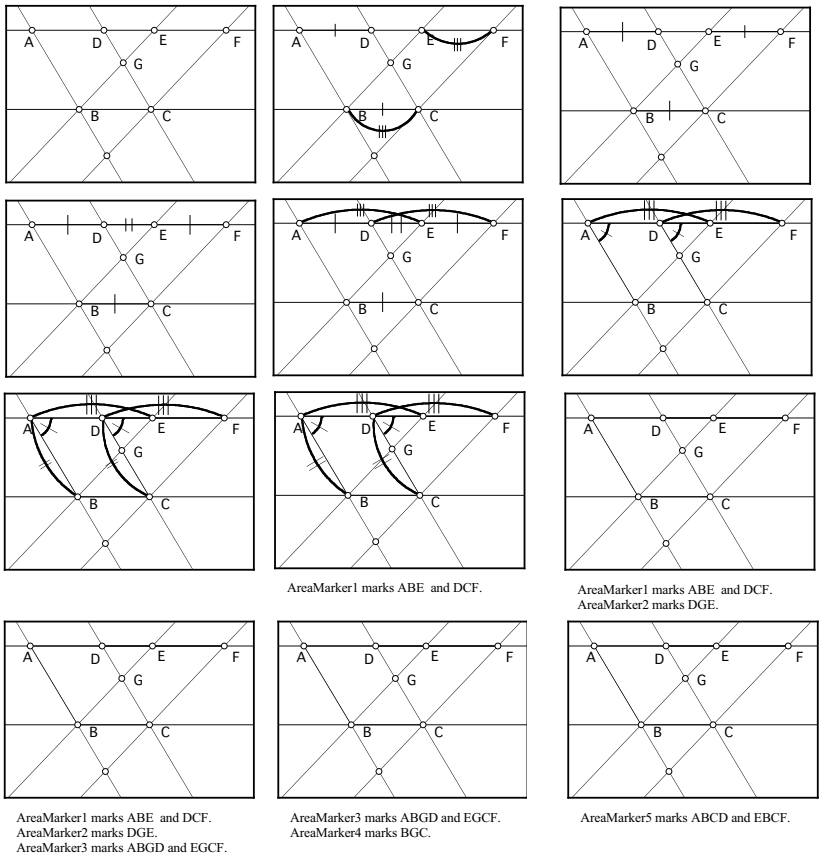


FIGURE 15 The proof of Book I, Proposition 35 in **FG′**.

follows the seventh by SAS, and the ninth follows from the eighth by several applications of rule R6. The tenth follows from the ninth by the area subtraction rule R2.5 applied to the areas marked by AreaMarker1 and AreaMarker2. The eleventh follows from the tenth by R6, and the final diagram then follows by the addition rule R2 applied to the areas marked by AreaMarker3 and AreaMarker4.

Notice that in this proof, we have used previously proven theorems as new proof rules; this practice is validated and discussed in Section 4.1.

With the addition of these rules, **FG′** should be able to duplicate all of the proof methods employed in the first four books of Euclid's *Elements*.

It should be noted that the formal system **FG′** is strictly stronger than the formal system **FG**. This does come at a slight cost, however. Because **FG′** allows markings of objects not normally marked congruent in most informal treatments of geometry, it could be argued that the diagrams of **FG** are better formal representatives of informal geometric diagrams than those of **FG′**. In any case, the reader should be warned that several of the results in the rest of this book refer specifically to **FG**, and may not apply to **FG′**. In particular, **CDEG**, discussed in Section 3.5 and Appendix D, is an implementation of part of **FG**; it doesn't allow the marking of di-arcs or dregions, and also does not implement **FG**'s transformation rules. Also, one of the complexity-theoretic results proven in Miller (2006) and discussed in Section 4.2, which says that it can be decided in polynomial space whether or not a diagram is satisfiable, may not hold for **FG′**; for details, see Section 4.2.

3.5 CDEG

In Section 1.2, it was pointed out that a proof system is completely formal if it can be implemented on a computer. Chapter 2 gave a careful formal definition of a diagram, and this chapter has given rules for manipulating such diagrams. However, given the complexity of the definitions, a skeptic might not be sure that this system is completely formal until seeing a computer implementation. This is especially true given the fact that another diagrammatic formal system which we now know to be unsound, Isabel Luengo's **DS1**, discussed in Appendix C, was published and studied by a number of smart people before its flaws were found.

To help convince such a skeptic, the construction and inference rules of **FG** have in fact been implemented in the computer system **CDEG**. In this section, we give a flavor of how this system works and how it implements these rules.

First of all, let's look at how **CDEG** implements our definition of what a diagram is. There were two pieces to this definition: in Section 2.1 we defined a diagram to be a certain kind of geometric object, and then in Section 2.2, we refined this definition to an algebraic structure that captures just the topological structure of a diagram, which is the part of the diagram that we want to consider to be information-bearing. **CDEG** has two ways of representing diagrams, which are analogous to these two parts of our definition. Internally, **CDEG** represents diagrams using a data structure that is based on the second, algebraic definition of a diagram graph structure. Its internal computations use this data structure exclusively. However, when the program outputs diagrams for human users to see, it offers two different choices. One way that it can display a diagram is as a string of text which lists all of the information found in the data structure; the other is as an actual geometric diagram with the given topological structure.

As a simple example, let's consider the first diagram in Figure 1, which contains a single line segment. We can create this diagram in **CDEG** by starting with the empty privative diagram, adding two new points, and then telling **CDEG** to connect them. If we then ask **CDEG** for its text description of this diagram, it prints the following:

```
Diagram #1:
dot13 is surrounded by: region4 solid15
dot12 is surrounded by: region4 solid15
solid15 ends at dots dot12 and dot13
frame3 ends at loop in regions region4 and outerregion
dline14 is made up of dot13 solid15 dot12

region4 has boundry: frame3
        and contents:
Component #1: dot13 solid15 dot12 solid15
```

If we then ask **CDEG** to display the graphical version of this diagram, it produces the picture shown in Figure 16.[15]

We can compare this description with the definition of a diagram graph structure given in Definition 6. Each piece of the diagram is assigned a unique number that identifies it, along with a name that

[15] All of the **CDEG** diagrams included in this book have been reproduced as they were output by the **CDEG** program, but with two minor modifications: (1) the colors that **CDEG** uses to identify different lines and circles have been changed to different shades of gray; and (2), the locations of the numbers labeling the different segments (such as that of the number *15* in Figure 16) have been altered in order to make the numbers more legible, since **CDEG** often places them so that they are obscured by the dots at the ends of the segments.

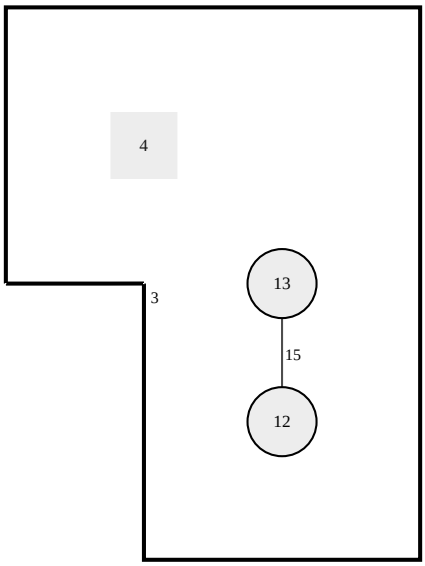


FIGURE 16 A **CDEG** diagram showing a single line segment.

identifies what kind of object it is. Each of the following clauses identifies which pieces of this diagram satisfy the corresponding clause of Definition 6.

1. The diagram contains two vertices, called `dot12` and `dot13`.
2. It contains two edges, `solid15` and `frame3`.
3. Each of the two vertices has only a single edge connected to it, `solid15`. Note that the data structure actually also includes information about what regions occur between the edges coming out from a vertex. In general, it would list all of the edges and regions surrounding the vertex, in clockwise order. The information about the regions is deducible from the other information in the diagram graph structure, but is included here for ease of computation.
4. In order to specify the two-dimensional cell-complexes, we just need to which vertices and edges occur around the boundary of each region. In this case, there is only one region, surrounded entirely by the frame. Note that we are allowing our graph to contain looped edges, like the frame in this diagram, that don't contain any vertices; and we are considering each such loop be a doubly connected component of the graph for the purposes of our definition.
5. The diagram contains a single non-outermost connected component, which contains `dot13`, solid15, and dot12, and which lies in `region4`. Again note that, for ease of computation, the data structure lists the whole outer boundary of the connected component in clockwise order. Thus, `solid15` is listed twice in the list giving `Component #1` of `region4`, because it occurs twice as we traverse the boundary of this component in clockwise order.
6. The diagram contains no pseudo-dots, because no line intersects the frame, so every vertex in the graph is a dot, and is identified as such in its name. Thus, $\mathrm{DOTS}(S) = \{$`dot12`, `dot13`$\}$.
7. The elements of the sets $\mathrm{SOLID}(S)$ and $\mathrm{DOTTED}(S)$ are identified by their names; in this case, $\mathrm{DOTTED}(S)$ is empty because the diagram doesn't contain any circles, while $\mathrm{SOLID}(S)$ contains a single element, `solid15`.
8. $\mathrm{SL}(S)$ contains a single subset of $E(S)$ because the pieces of the diagram are only part of one straight line, namely `dline14`. The subset of $E(S)$ that corresponds to this line is $\{$`solid15`$\}$, since that is the only segment on this line. Note, however, that in the data structure, in addition to the segment, the description of the

dline also gives the dots that are part of it, again for ease of computation.

9. Finally, the set CIRC(S) is empty because there are no circles in the diagram.

As a more complicated example, we can consider the fifth diagram shown in Figure 1. **CDEG**'s graphical version of this diagram is shown in Figure 17, and if we print out this much more complicated diagram, it looks like this:

```
CDEG(1/1)% p
Diagram #1:
dot433 is surrounded by: dottedseg434 region437
     dottedseg436 region438 dottedseg435 region466
     dottedseg468 region467
dot289 is surrounded by: dottedseg290 region438
     dottedseg436 region437 dottedseg434 region467
     dottedseg294 region1461 solid1462 region1762
     solid1763 region1761
dot13 is surrounded by: region466 dottedseg435 region438
     dottedseg290 region1761 solid1763 region1762 solid15
dot12 is surrounded by: region1762 solid1462 region1461
     dottedseg294 region467 dottedseg468 region466 solid15
solid1763 ends at dots dot13 and dot289
solid1462 ends at dots dot12 and dot289
dottedseg468 ends at dots dot433 and dot12
dottedseg436 ends at dots dot289 and dot433
dottedseg434 ends at dots dot289 and dot433
dottedseg435 ends at dots dot433 and dot13
dottedseg294 ends at dots dot12 and dot289
dottedseg290 ends at dots dot13 and dot289
solid15 ends at dots dot12 and dot13
frame3 ends at loop in regions region437 and outerregion
dline1463 is made up of dot289 solid1763 dot13
dline484 is made up of dot289 solid1462 dot12
dline14 is made up of dot13 solid15 dot12
circle87 has center dot13 and boundry dottedseg468 dot433
     dottedseg436 dot289 dottedseg294 dot12
circle23 has center dot12 and boundry dottedseg290 dot289
     dottedseg434 dot433 dottedseg435 dot13

region1761 has boundry: solid1763 dot13 dottedseg290
     dot289
```

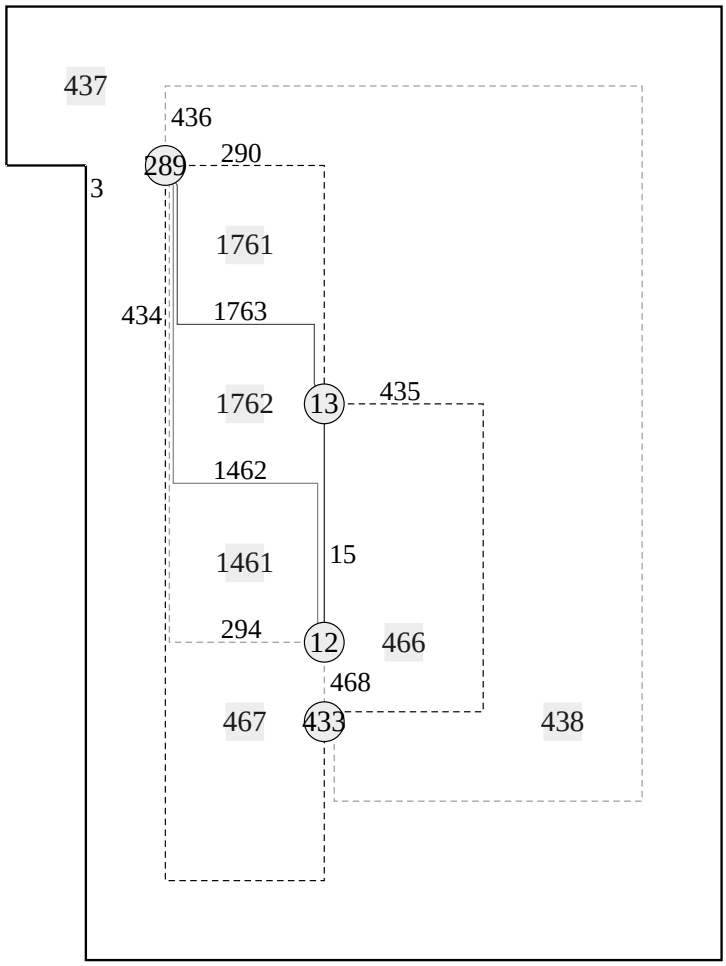


FIGURE 17 A **CDEG** diagram showing the triangle obtained in the proof of Euclid's First Proposition.

```
        and contents:

region1762 has boundry: solid1763 dot289 solid1462 dot12
     solid15 dot13
        and contents:

region1461 has boundry: solid1462 dot289 dottedseg294
     dot12
        and contents:

region466 has boundry: dottedseg468 dot433 dottedseg435
     dot13 solid15 dot12
        and contents:

region467 has boundry: dottedseg468 dot12 dottedseg294
     dot289 dottedseg434 dot433
        and contents:

region438 has boundry: dottedseg436 dot289 dottedseg290
     dot13 dottedseg435 dot433
        and contents:

region437 has boundry: frame3
        and contents:
Component #1: dottedseg436 dot433 dottedseg434 dot289
```

The graphical version of this diagram may seem much more complicated than the version created by hand and shown in Figure 1, but, in fact, the two diagrams are equivalent, and the diagram produced by the computer is still much easier to decipher than the printout, which also contains the same information. It should also be noted that in practice, the diagram produced by the computer is somewhat easier to read than the black and white version printed here, because each dline and dcircle is assigned a unique color which makes it much easier to tell which segments belong to the same object. As in the version created by hand, the segments that are parts of circles are dotted, while the the three segments that make up the triangle are drawn as solid lines.

The graphical version of the diagram looks significantly different than the diagram that a human being would produce because it is laid out by a standard graph drawing package that takes as an input a purely topological description of the graph. It has no knowledge of the meaning of any of the pieces of the graph. On the other hand, having

a representation like this is useful because it makes it perfectly clear that the diagram is merely a syntactic representation standing in for the configuration of geometric objects that it represents.

Looking next at the printed description of the diagram, we see again that

1. Each dot lists the segments and regions that surround it, in clockwise order, although the definition of a diagram graph structure only requires that it list the segments;
2. Each segment lists its ends, although the definition of a diagram graph structure doesn't require this, as the ends of a segment can be deduced from information given for each vertex;
3. Each dline lists the dots and segments that lie on it, in order, although the definition only requires that it list the segments, as the dots could again be deduced;
4. Each circle lists its center and each dot and segment that lies on it, in order, although, as before, it isn't required to list the dots;
5. Each region lists its boundary in clockwise order, which is enough information to construct the two-dimensional cell complex described in the definition; and
6. Each region lists the connected components that lie inside it, in the form of a clockwise listing of the outside boundary of the connected component. In this case, there is again only one non-outermost connected component, in `region437`. Although there are quite a few dots and segments in this component, its boundary only contains the two dots and two dotted segments listed.

All of the extra information in **CDEG**'s internal representation of a diagram is included to make it easier for it to compute the result of applying construction rules to the diagram. As an example of this, let's consider the case analysis problem from Section 3.1 illustrated in Figures 8 and 9. We can duplicate Figure 8 in **CDEG** by starting with the diagram shown in Figure 16 and adding two more dots to `region4`, which **CDEG** labels as `dot23` and `dot24`, producing the diagram shown in Figure 18. We now want to consider the means by which **CDEG** figures out how to produce all of the cases that can occur when a segment is drawn from `dot23` to `dot24`.

CDEG knows that the new segment must start at `dot23`. Because the only part of the diagram listed as surrounding this dot is `region4`, the segment must go from `dot23` to `region4`. From `region4`, the segment must next hit either a part of the region's boundary, or else part of the boundary of one of the connected components contained in the region. In this case, the boundary of `region4` is the frame, and dsegs

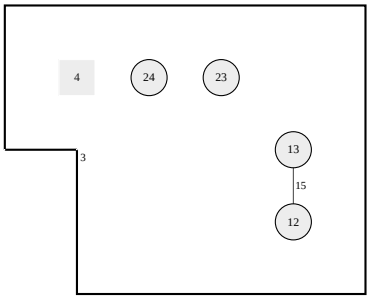


FIGURE 18 A **CDEG** diagram corresponding to the diagram shown in Figure 8.

aren't allowed to hit the frame, so that case is eliminated. Furthermore, the dseg isn't allowed to loop back to the dot it came from. This means that the segment must next hit one of the other two possible connected components in `region4`: the one containing `dot24`, or the one containing `dot12`, `dot13`, and `solid15`. If it hits the first one of these connected components, then the new segment has reached its destination and we are finished, left with the upper left hand diagram in Figure 19. If it instead hits the other component, it can hit it anywhere along the boundary of that component. That is why, in the printed description of the diagram, the component is identified by listing this whole boundary as follows: `Component #1: dot13 solid15 dot12 solid15`. Notice that `solid15` occurs twice on this list, and this is exactly what we want, because the new dseg can hit `solid15` from either side, giving us the two different cases shown in the two right hand diagrams in Figure 19. In each of these two cases, the segment must continue through to the other side after hitting `solid15` (because of the rules for avoiding dtangency of different straight lines), back into `region4`, and then finally to `dot24`, because every other place it could go from there would violate some condition for well-formedness. (**CDEG**, of course, generates all of the possibilities at each stage, and then discards those that violate some rule.) In the next case, if the new segment hits dot13, then there are three different ways for it to continue: along `solid15`, or back into `region4` on either side of it. These three cases are shown in the two topmost diagrams in Figure 20 and the bottom left-hand diagram in Figure 19. Finally, the new segment could hit `dot12` first, giving three more cases shown in the remaining three diagrams in Figure 20. Altogether, we get the same nine cases shown in Figure 9.

For readers who are interested, a complete transcript of a **CDEG**

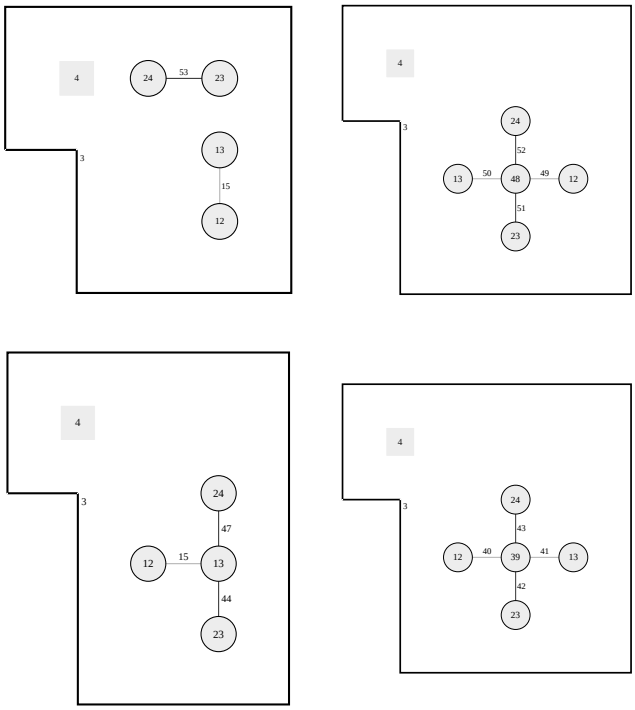


FIGURE 19 Four of the **CDEG** diagrams corresponding to those in Figure 9.

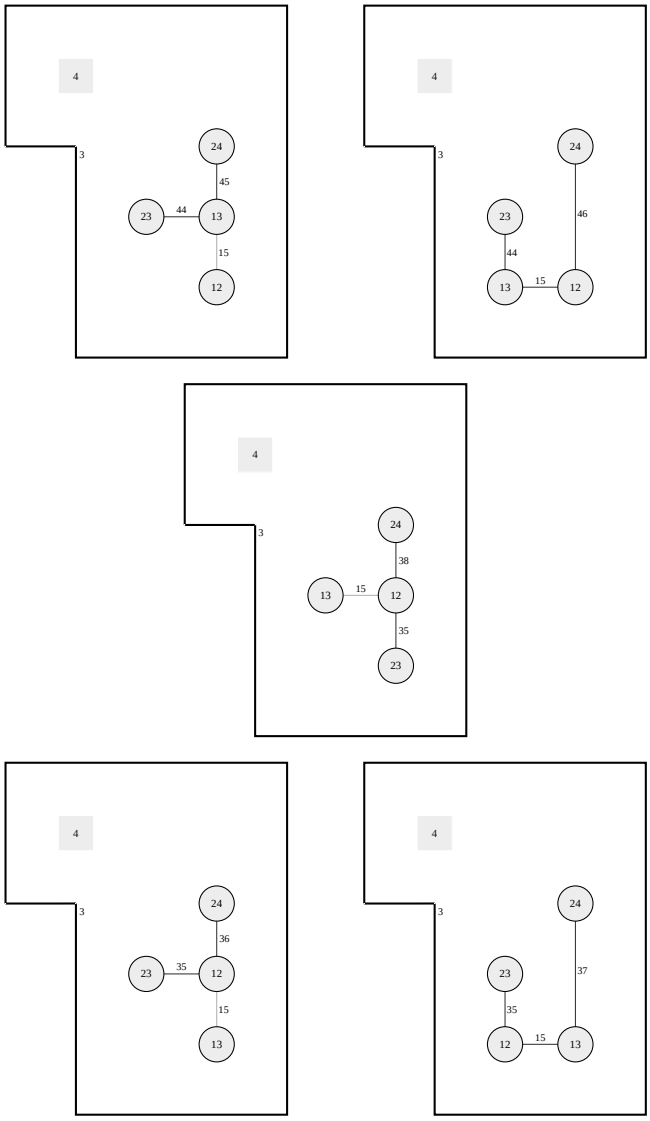


FIGURE 20 Five of the **CDEG** diagrams corresponding to those in Figure 9.

session that produced all of these diagrams is included in Appendix D. This transcript shows how the complete proof of Euclid's Proposition 1 can be duplicated in **CDEG**.

4

Meta-mathematical Results

In this chapter, we will look at ways that we can use the formal system defined in the previous two chapters to shed light on proof practices in geometry and on the logical structure of geometry.

4.1 Lemma Incorporation

One of the largest benefits of formalizing our proof system is that the formal results can shed light on proof practices in the informal system. For example, it is very common for informal proofs of geometric facts to rely on other theorems that have already been proven. This use of previously proved lemmas in proofs is a normal practice in most of mathematics, but it is particularly common in geometry, and while it is clear how lemma incorporation works formally in traditional sentential proofs, it is less clear how it should work in diagrammatic proofs. In a sentential proof, whenever a lemma is used, the proof of the lemma can be inserted to give a proof that doesn't rely on the lemma. Furthermore, using the lemma doesn't shorten the proof at all: the length of the new proof is the same as the length of the original proof plus the length of the proof of the lemma.

In a diagrammatic proof, on the other hand, the diagram in which you want to apply a lemma is generally much more complicated than the diagram in which you first proved the lemma. This makes it less clear how lemma incorporation should work formally, but points towards one of the reasons that it is useful: we can save ourselves some work by proving lemmas in the simplest possible environments. In fact, some diagrammatic proofs can be made exponentially shorter by using lemmas. This is because lemma incorporation can partially make up for one of the great weaknesses of diagrams: unnecessary case analysis stemming from too much information in the diagram.

For example, consider a diagram like the first one in Figure 21 that

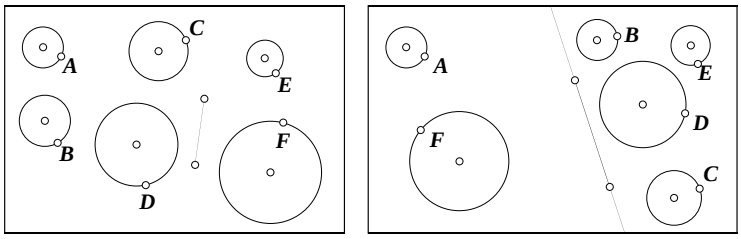


FIGURE 21 Extending a line can give rise to exponentially many new cases.

contains n circles and a line segment. What happens when we try to extend this line segment to a line? Each circle could end up on either side of the line (or it could be on the line), so there are at least 2^n different cases which could result. One of these is shown by the second diagram in Figure 21. This is a real disadvantage of working diagrammatically: applying one construction rule can give us exponentially many new cases to consider. What's worse, the extra cases might be totally superfluous—what if the line segment was being extended for some part of the proof that had nothing to do with the circles? In that case, one of the strengths of this system—that it can do complete case analysis of the diagrams—becomes a weakness, because we are forced to consider cases that may never play any important role in a proof.

This is where lemma incorporation becomes useful. By proving lemmas in a simple environment and then applying them, we can avoid unnecessary case analysis. As an example, consider what happens if we want to apply SAS to a diagram that contains n disconnected components along with the two triangles. If we apply SAS as a lemma, we can conclude that the triangles are congruent in one step, so that the total length of the proof is the same regardless of how big n is, even if we include the length of the proof of the lemma in the length of the proof. But if we try to mimic the proof of the lemma directly in the diagram with the n extra disconnected components, we will get at least 2^n extra cases, because the proof involves moving one triangle to the other, and each of the disconnected components might or might not end up lying inside the moved triangle. As we have mentioned, commentators often point out that Euclid proved SAS using superposition and then used SAS rather than superposition to prove his other theorems, and sometimes claim that this indicates that he viewed superposition as being a suspect method of proof. (See for example Thomas Heath's commentary on Euclid's Fourth Proposition.[16]) Using SAS as a lemma

[16](Euclid, 1956, pp. 224–231, pp. 249-250)

makes proofs simpler and shorter than the corresponding proofs that use superposition directly, however, so we shouldn't be surprised that this is what Euclid did.

So how does lemma incorporation work in the diagrammatic world? We would like to claim that, just as in the sentential world, we can incorporate lemmas into our proofs in such a way that any diagram that we can derive from another using a lemma could have been derived without using the lemma. First, we need a way to combine the diagram from the lemma with the original diagram.

Definition 10 The ***unification*** of primitive diagrams a and b (written as $\text{unif}(a, b)$) is the diagram array containing all minimal diagrams d that contain both a and b as subdiagrams.

Recall that a primitive diagram a is a subdiagram of d if a can be obtained from d by using rule C4 to erase pieces of d. We say that a segment in d ***comes from*** a if it is part of a set of segments in d that is related by the counterpart relation to a set of segments in a, and that a dot in d comes from a if any of the segments that it intersects come from a. Notice that a diagram d is in $\text{unif}(a, b)$ just if it contains a and b as subdiagrams, and every dot and segment in d comes from either a or b.

Next, we'd like to extend these ideas to diagram arrays. If A and B are diagram arrays, we say that A is a ***subdiagram*** of B via the matching function $f : B \to A$ iff for every primitive diagram b in B, $f(b)$ is a primitive diagram a in A such that a is a subdiagram of b. Likewise, if 2^A denotes the set of subsets of the set of diagrams in A, then given two diagram arrays A and B and a function $g : B \to 2^A$, the ***unification*** $\text{Unif}_g(A, B)$ of A and B with respect to g is the smallest diagram array such that for each primitive diagram b in B and each primitive diagram a in $g(b)$, $\text{Unif}_g(A, B)$ contains all the diagrams in $\text{unif}(a, b)$. Finally, if $A \vdash A^*$ via derivation D, then we can define an ***ancestor relation*** ancestor_D between primitive diagrams in A and primitive diagrams in A^* in the natural way: every primitive diagram in a diagram array at one step of the derivation is descended from a unique primitive diagram in the preceding array, unless rule R4 removing copies is applied, in which case the remaining copy has two ancestors in the preceding generation. Thus, ancestor_D relates $a_0 \in A$ and $a_1 \in A^*$ if a_1 is descended from a_0.

As an example of how this works, consider Figure 22. Let B be the diagram array on the first line, let A be the diagram array on the second line, and let A^* be the diagram array on the last line. A is a subdiagram of B via the matching function f shown by the arrows between the first

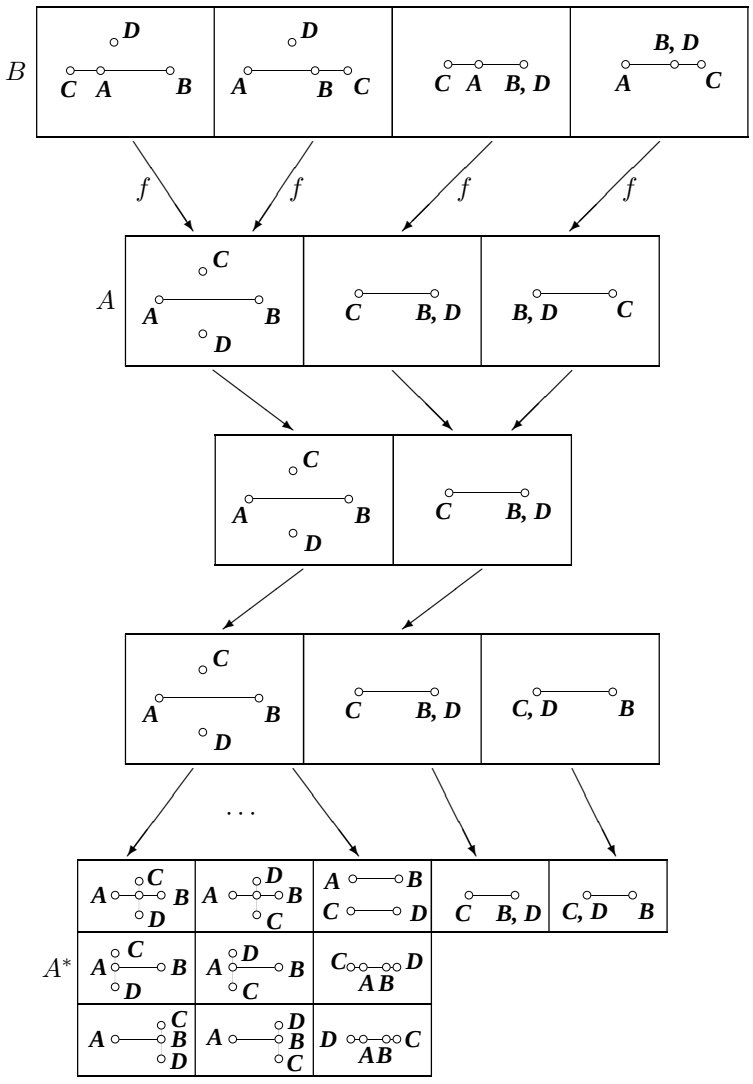


FIGURE 22 An example of lemma incorporation.

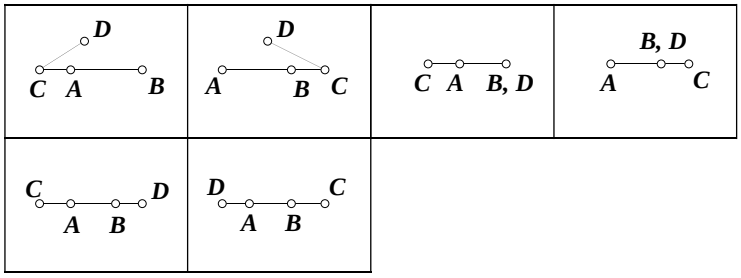


FIGURE 23 The result of unifying B and A^* in Figure 22.

and second lines. A^* can be derived from A in three steps: first, the third diagram in A is removed using rule R4, since it is a copy of the second diagram in A; next, a new diagram is added to the array using rule C5; and finally, rule C1 is applied to points C and D in the first diagram in the array. This yields A^*, in which the first nine primitive diagrams correspond to the nine different ways of connecting points C and D, and the last two primitive diagrams are the same as in the preceding array. The arrows in Figure 22 show how the diagrams are descended from the diagrams in the preceding array at each step. The first nine diagrams in A^* are descended from the first diagram in A; the tenth diagram is descended from both of the other diagrams in A; and the eleventh diagram in A^* isn't descended from any diagram in A.

How would we apply this derivation to B? Intuitively, it seems that we would want to unify each diagram in B with the diagrams that are descended from the corresponding diagrams in A. This would give us the unification of A^* and B with respect to the function g that takes each diagram b in B to the set of descendants of $f(b)$. In this particular case, we would unify each of the first two diagrams in B with each of the first nine diagrams in A^*, and we would unify the other two diagrams in B with the tenth diagram in A^*. Doing this gives us the diagram array B^* shown in Figure 23.

In this case, it is clear that B^* can be derived from B directly, by applying rule C1 to points C and D in the first two diagrams in B. We would like to know that any diagram like this, the result of unifying a diagram B with a diagram derived from a subdiagram of B, could have been derived directly from B. This is what the Lemma Incorporation Theorem tells us.

Theorem 1 (Lemma Incorporation) *Assume that A is a subdiagram of B via matching function f, and $A \vdash A^*$ via derivation D. Let*

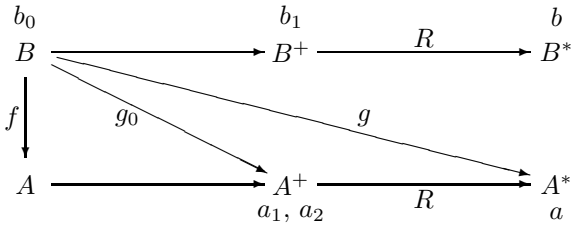


FIGURE 24 Lemma Incorporation.

g and U be defined so that $g(b) = \{d \in A^* | ancestor_D(f(b), d)\}$ and $U = Unif_g(A^*, B)$. Then $B \vdash U$.

Proof. We prove this by induction on the length of the proof that $A \vdash A^*$.

(Base Case) In this case, $A = A^*$, and ancestor_D is the identity function, so $\text{Unif}_g(A^*, B) = \text{Unif}_f(A, B) = B$, since A is a subdiagram of B via f. So $B \vdash \text{Unif}_g(A^*, B) = B$ trivially.

(Inductive Case) Assume that $A \vdash A^+$, that $A^+ \vdash A^*$ in one step via construction, inference, or transformation rule R, and that $\text{Unif}_{g_0}(A^+, B)$ has already been derived from B, by the inductive hypothesis, where $g_0(b) = \{d \in A^+ | \text{ancestor}_D(f(b), d)\}$. Let $U = \text{Unif}_g(A^*, B)$, $B^+ = \text{Unif}_{g_0}(A^+, B)$, and let B^* denote the result of applying rule R to the primitive diagrams in B^+ that were obtained as unifications of the primitive diagram in A^+ to which R was applied, unless R is C5 or R4. The resulting situation is illustrated in Figure 24. We apply the rule R to the same objects in B^+ that it was applied to in A^+; this is possible since A^+ is a subdiagram of B^+. Since $B \vdash B^*$ (because $B \vdash B^+$ by the inductive hypothesis and $B^+ \vdash B^*$ by rule R), it suffices to show that $B^* \subseteq U$, since then $B^* \vdash U$ by rule C5. We will show this for all of the rules except for C5 and R4, in which cases we will show that $U = B^+$, so that U is again derivable.

- *R is a construction, inference, or transformation rule adding object(s) l to diagram $a_1 \in A^+$.* Here, l can consist of dots, segments, and/or markers. We want to show that $B^* \subseteq U$. Pick $b \in B^*$. If b wasn't modified by R, then $b \in B^+$, and b corresponds to a diagram $a_2 \in A^+$ other than a_1 such that $b \in \text{unif}(a_2, b_0)$ for some diagram $b_0 \in B$. Since $a_2 \neq a_1$, $a_2 \in A^*$, so $b \in \text{unif}(a_2, b_0) \subseteq \text{Unif}_g(A^*, B) = U$. On the other hand, if b was modified by R, then b is a descendent of some diagram $b_1 \in B^+$, where $b_1 \in \text{unif}(a_1, b_0)$ for some $b_0 \in B$. This means that b_1 contains a_1 as a subdiagram, and the pieces of b_1 that don't come from a_1 must come from b_0. Since b is b_1 with

added object l, it contains a subdiagram a that consists of a_1 with added object l. This subdiagram is therefore in A^*. Thus, since all of the pieces that are in b_1 that don't come from a_1 come from b_0, all of the pieces of b that don't come from a come from b_0 also, since a and b consist of a_1 and b_1 with l added in the same way. So b is in $\text{unif}(a, b_0)$ and therefore in U; so $B^* \subseteq U$.

- *R is a rule (either C4 or the second part of R6) that removes some object(s) l from the diagram.* To show $B^* \subseteq U$, assume $b \in B^*$ is descended from $b_1 \in B^+$ and that $b_1 \in \text{unif}(a_1, b_0)$, where $a_1 \in A^+$ and $b_0 \in B$. Either a_1 wasn't modified by R, in which case $b \in U$ as in the previous case, or else there exists a diagram $a \in A^*$ that consists of a_1 with l removed. In this case, $b \in \text{unif}(a, b_0)$. We know this because any diagrammatic object *do* in b is in $b_1 = \text{unif}(a_1, b_0)$ and not in l, so it is in either a_1 or b_0, and isn't in l, so if it's in a_1, then it is in fact in a. So *do* is either in a or in b_0, so it's in $\text{unif}(a, b_0)$. So $B^* \subseteq U$.
- *R is a rule (either R5a, or R5b) that removes an inconsistent diagram D from A^+.* Pick $b \in B^*$. Then b was also in B^+. (Because some other diagram was removed.) So $b \in \text{unif}(a_1, b_0)$ for some $a_1 \in A^+$ and $b_0 \in B$, and a_1 wasn't the diagram that was removed from A^+ (since then b would have been removed from B^+). It therefore follows that $a_1 \in A^*$, and so $b \in \text{Unif}_g(A^*, B)$.
- *R is rule C5, adding a new diagram n to diagram array A^+.* Then, since n isn't in the image of g, $\text{Unif}_g(A^*, B) = \text{Unif}_{g_0}(A^+, B) = B^+$. So $U = B^+$ and U has therefore already been proven in this case.
- *R is rule R4, removing a diagram d_2 that is a copy of some other diagram d_1 in A^+.* We want to show that in this case, $B^+ = U$, so that U has already been derived from B. First, we show that $B^+ \subseteq U$. Choose $b_1 \in B^+$. We know that $b_1 \in \text{unif}(b_0, a_1)$ for some $b_0 \in B$ and $a_1 \in A^+$. If $a_1 \neq d_2$, then a_1 is still in A^*, so $b_1 \in \text{unif}(a_1, b_0) \subseteq \text{Unif}_g(A^*, B) = U$ as required. On the other hand, if $a_1 = d_2$, this means that $g_0(b_0)$ contains d_2, in which case $g(b_0)$ contains d_1, since d_1 is an ancestor of d_2. Since d_1 and d_2 are copies of one another, $\text{unif}(b_0, d_2) = \text{unif}(b_0, d_1)$. So, in particular, $b_1 \in \text{unif}(b_0, d_2) = \text{unif}(b_0, d_1) \subseteq \text{Unif}_g(A^*, B) = U$, as required. So $B^+ \subseteq U$.

  On the other hand, if $b \in U$, then $b \in \text{unif}(b_0, a)$ for some $b_0 \in B$ and $a \in A^*$ such that $a \in g(b_0)$. Then either $a \in g_0(b_0)$, in which case $b \in \text{unif}(b_0, a) \subseteq \text{Unif}_{g_0}(A^+, B) = B^+$, or else $a = d_1$ and $d_2 \in g_0(b_0)$, in which case $b \in \text{unif}(b_0, a) = \text{unif}(b_0, d_1) = \text{unif}(b_0, d_2) \subseteq \text{Unif}_{g_0}(A^+, B) = B^+$. So $b \in B^+$; so $U \subseteq B^+$; and so $U = B^+$, which

is what we were trying to show.

So U is derivable in each case, as required. □

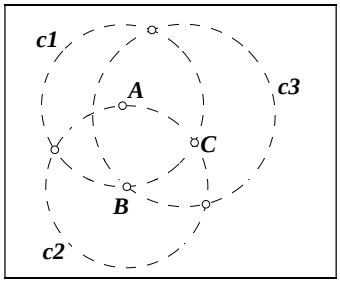
4.2 Satisfiable and Unsatisfiable Diagrams

Another benefit of having formalized our proof system is that we can then use technical methods of analyzing formal systems to better understand the properties of our proofs. In this section, we will look at what complexity theory can tell us about our system, and the practical implications of these results. Complexity theory is the mathematical theory of how difficult or time consuming different kinds of computations are.

In Section 2.3, we showed that only nicely well-formed primitive diagrams have models. This isn't surprising, since the definition of a nicely well-formed primitive diagram was designed to eliminate diagrams that represented unrealizable situations. A more interesting question might be: how well did our definitions succeed at eliminating these unrealizable situations? That is: did we succeed in eliminating *all* of the unsatisfiable diagrams, or are there still some nwfpds with no models?

Unfortunately, the answer is that there are indeed unsatisfiable nwfpds. First of all, if we allow marked diagrams, it is easy to find examples of unsatisfiable diagrams. For example, any diagram that can be eliminated using inference rules R5a and R5b (CS and CA) is unsatisfiable. Another example is given by a diagram that contains a circle with a radius that is marked congruent to a (different) diameter; and there are many others. However, even unmarked diagrams may not be satisfiable. Figure 25 gives an example of such a nwfpd that is unsatisfiable: it follows from the diagram that $\overline{AB} \cong \overline{AC}$, since both segments are radii of $c1$, and similarly $\overline{AB} \cong \overline{BC}$, because both of these segments are radii of $c2$; so by transitivity, we should have $\overline{AC} \cong \overline{BC}$; but according to the diagram, B is on $c3$ and A isn't, so $\overline{AC} \ncong \overline{BC}$ and the diagram is unsatisfiable. The problem with this diagram seems to be that lengths have snuck in here via circles, so the diagram isn't just showing topological information; it's also showing geometric information about which line segments are the same length.

One might next hypothesize that any nwfpd that contains neither dcircles nor dlines that aren't proper has a model, but this isn't true either. Figure 26 shows such a nwfpd without a model. The easiest way to see that this diagram isn't satisfiable is to look at the rectangle in the center of the diagram. There are two crossing lines that go through the corners of this rectangle, and two other crossing lines that are copies of these that have been parallel transported downward along the sides of



Dot A is the center of dcircle c1;
dot B is the center of dcircle c2; and
dot C is the center of dcircle c3.

FIGURE 25 An unsatisfiable nwfpd.

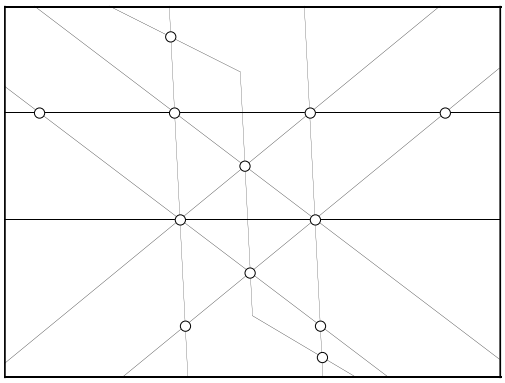


FIGURE 26 Another.

the rectangle from the upper corners to the lower corners. So the point of intersection of the crossing lines has also been moved down parallel to the sides of the rectangle. The line that goes through the original point of intersection and the transported point of intersection should therefore be parallel to the sides of the rectangle; but it intersects them, so the diagram isn't satisfiable. This shows that lengths can also sneak in via parallel lines.

Finally, one might hypothesize that at least any unmarked primitive diagram that doesn't contain any circles or parallel lines should be satisfiable, but even this isn't true. Figure 27 shows a nwfpd which only contains line segments, and which is nevertheless unsatisfiable because according to Desargues' theorem, in any model of this diagram, line XY would have to intersect line $B'C'$ at point Z. The usual proof of Desargues' theorem is as follows: imagine that the diagram shows a two-dimensional projection of a three-dimensional picture of a pyramid

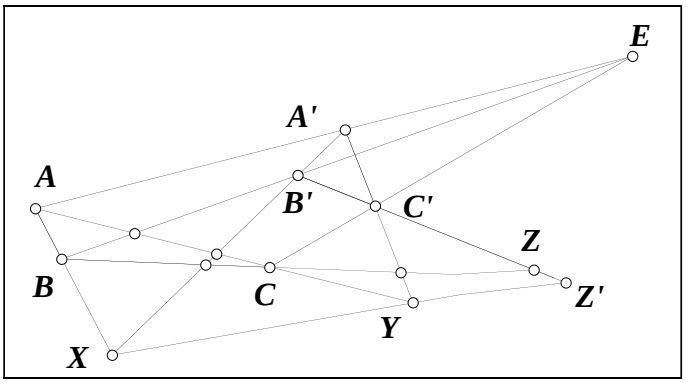


FIGURE 27 An unsatisfiable nwfpd containing nothing but unmarked dsegs.

with base ABC and summit vertex E. Then the triangles ABC and $A'B'C'$ determine two different planes P_1 and P_2 in three-space. Lines AB and $A'B'$ meet in three-space, because they both lie in the plane determined by triangle ABE, and since AB lies in P_1 and $A'B'$ lies in P_2, their point of intersection X must lie in the intersection of P_1 and P_2. Likewise, if Y is the intersection of AC and $A'C'$ and Z is the intersection of BC and $B'C'$, then Y and Z also lie in the intersection of the two planes. Since two planes intersect in a line, this means that X, Y, and Z should be collinear; but in the given diagram, point Z doesn't fall on line the line XY, so the diagram is unsatisfiable.

The preceding examples show that our definition of what it means to be nicely well-formed is still too broad, because there are diagrams that are nicely well-formed but are still unsatisfiable. An obvious next question is: is there some additional set of conditions that can be added to those for nice well-formedness that will eliminate all of the unsatisfiable diagrams? It would be extremely convenient to find such a set of conditions. For example, consider what happens when we apply one of the construction rules to a satisfiable diagram. We get back an array of possible results. As it now stands, we know that because the construction rules are sound, at least one of the diagrams that we get back must be satisfiable, but many of them may not be satisfiable. If we could find a set of conditions that eliminated these unsatisfiable diagrams, then we wouldn't have to waste our time looking at these extra cases. So such a set of conditions would be extremely powerful.

The very fact that such a set of conditions would be so powerful might make us suspect they would be *too* powerful, and that such a set of conditions is impossible to find. But somewhat surprisingly, it is in

fact possible to compute whether or not a given diagram is satisfiable, at least for diagrams in **FG**. In Miller (2006), it is shown how to translate our definition of satisfiability in **FG** into the first-order language of real arithmetic, so that given a diagram, we can find a corresponding sentence that is true if and only if the given diagram is satisfiable.[17] Once we have this formula, we can then apply Tarski's Theorem, which says that there is a procedure for deciding if a given sentence of the first-order language of arithmetic is true or false (as a statement about the real numbers). (See Tarski (1951).) This means that we could define a primitive diagram to be ***strongly nicely well-formed*** if it is nicely well-formed and Tarski's decision procedure says that it is satisfiable.[18] Then the strongly nicely well-formed diagrams would exactly capture the possible configurations of real Euclidean planes. The problem with this approach is that the decision procedure given by Tarski's theorem can take intractably long to run. (The formula derived in Miller (2006) that translates our definition of diagram satisfaction contains only existential quantifiers, and is therefore a Σ_1 formula. The best known general algorithm for deciding Σ_1 formulas over the real numbers is in the complexity class PSPACE, which contains algorithms that require polynomial computation space (memory) to run. In general, such algorithms are considered to be practically intractable because they can take exponentially long to evaluate.) A set of conditions that correctly determine if a diagram is satisfiable but take exponentially long to evaluate aren't really useful.

On the other hand, the fact that that particular method of deciding if a diagram is satisfiable can take intractably long doesn't necessarily mean that every such method will be intractable. So a new question is: is there a procedure that determines whether or not a given diagram is satisfiable in a reasonable amount of time? A procedure is usually considered to be tractable in the real world only if it runs in polynomial time. So, to be more specific, our question becomes: is there a polynomial-time algorithm for determining whether or not a given diagram is satisfiable? It turns out that the answer to this question is no, assuming that P $\neq$ NP as is widely believed to be the case. In Miller (2006) it is also shown that the diagram satisfiability problem is NP-hard, which means that no set of conditions that can be evaluated

[17] This is proven in Miller (2006) only for **FG**. Whether or not it can be extended to **FG′** is an open problem.

[18] If we are working in **FG′**, we could define it to be nicely well-formed if the diagram that results from erasing any markings of dregions or di-arcs is nicely well-formed in **FG**.

in polynomial time can determine if a given diagram is satisfiable.[19]

What does this mean? It means that, while diagram satisfiability can theoretically be decided by a computer, and case analysis in geometry can therefore be theoretically be done by a computer exactly in a way that will never return unsatisfiable diagrams, it is nonetheless impractical to do so, and any real world computer system or formal system will therefore sometimes unavoidably return extra unsatisfiable cases. This is perhaps not surprising given the examples above, but it does show us that the long tradition of considering and disposing of extra cases in geometry is unavoidable in a diagrammatic approach.

4.3 Transformations and Weaker Systems

A third benefit of having formalized our proof system is that we can now study the logical structure of that proof system.

Most formal systems for doing geometry (Hilbert's, for example) don't contain rules for doing symmetry transformations; rather, they include a version of the rule of inference SAS. In **FG**, SAS is a derived rule that we know can be proven in the same way Euclid proved his fourth proposition.

However, in our system, we can also consider sets of rules that are weaker than **FG**, so that A can no longer be derived from them, but which are still strong enough to prove some of the things that are normally proved using SAS. For example, consider the system **GS** ("Geometry of Segments") in which we have all the rules of construction, transformation, and inference except for CA. SAS is not a derived rule of this system. To see this, consider a modified definition of satisfaction in which $M \models D$ iff M's canonical unmarked diagram is equivalent to D's underlying unmarked diagram, and if two dsegs in D are marked with the same marker, then the corresponding dsegs in M's canonical marked diagram are also marked with the same marker (so that we have dropped the corresponding requirement for di-angles). All of the rules of **GS** are still sound with respect to the new notion of satisfaction (call it **GS**-satisfaction), but CA and SAS are no longer sound. This is because the definition of **GS**-satisfaction says that any two angles can be marked with the same marker even if they aren't really congruent,

[19] Thus, the diagram satisfiability problem is at least NP-hard, and could possibly be PSPACE-hard. In fact, in Miller (2006) it is also shown that that the diagram satisfiability problem has exactly the same complexity as the satisfiability problem for a particular subset of the Σ_1 formulas over the real numbers; and because this subset is quite expressive, it seems likely that the satisfaction problem for diagrams is of complexity close to that of the whole existential theory of the real numbers, for which the best known bound is that is is in PSPACE. See Renegar (1992) for a discussion of the complexity of the existential theory of the reals.

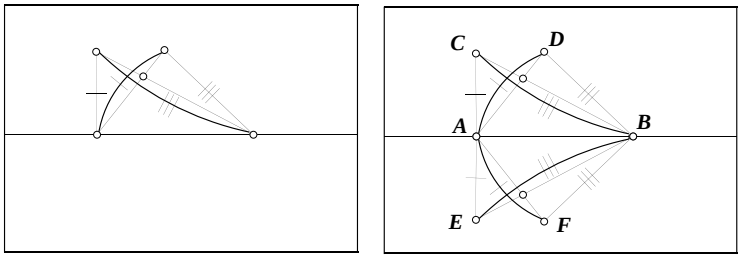


FIGURE 28 Steps in the proof of SSS.

so it is possible to have an angle properly contained in another with the same marking; and the corresponding angles of two triangles with congruent sides could be marked with the same marking even if they aren't really congruent, so that the resulting triangles aren't congruent either. Thus, neither CA nor SAS is derivable in **GS**, since it is impossible to derive an unsound rule from sound rules. On the other hand, many consequences of SAS still hold: for example, the SSS rule for triangle congruence can still be derived. (This is plausible, since the SSS rule is still sound with respect to **GS**-satisfiability.)

Here is a description of how to derive the SSS rule in the system **GS**: given two triangles whose sides are marked equivalent, use the symmetry transformations and CS to move the second triangle so that its first side coincides with the first side of the first triangle and the two triangles are oriented the same way. Either the second triangle lies precisely on top of the first, in which case we're done, or else we have a situation that looks like the first diagram in Figure 28. Reflect the two triangles over their common base line, giving the situation shown in the second diagram in Figure 28. Construct the circles c_1 and c_2 with centers A and B through point C. It follows from CS that if a circle is drawn with center Z through a point X and ZX is marked congruent to some other segment ZY also ending at Z, then the circle must also pass through Y; otherwise, if the circle intersects ray ZY at point W, then ZW can be marked congruent to ZX and therefore marked congruent to ZY, but one of ZY and ZW must be properly contained in the other, a contradiction by CS. The two circles c_1 and c_2 must therefore each intersect the four distinct points C, D, E, and F; but two distinct circles can only intersect in at most two points; a contradiction.

So what is the relationship between CA and SAS? They are in fact equivalent in **GS**. In the previous section, we showed how to derive SAS in **FG**; this shows that SAS can be derived from CA in **GS**. But

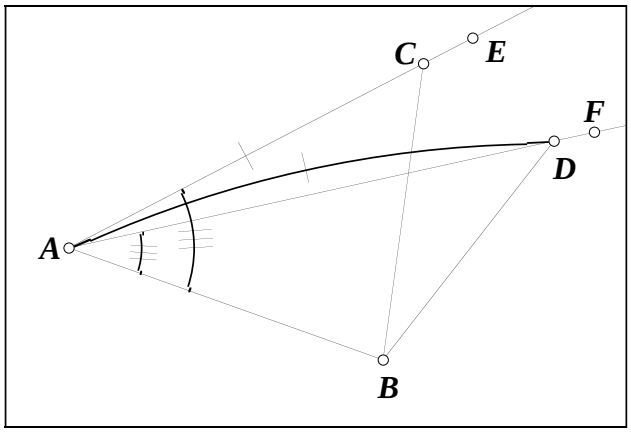


FIGURE 29 Deriving CA from SAS in **GS**.

CA can also be derived from SAS in **GS**, as follows. Let us be given a diagram in which two di-angles, one contained in the other, are marked with the same marking, as in Figure 29, and let us denote the di-angles BAE and BAF. We need to show how to eliminate this diagram in **GS**. To do this, we can mark off equal length segments AC and AD along AE and AF (using rule C0 to add a dot C along AE, using rule C3a to draw a circle about A through C, labeling the intersection of this circle with AF as D, and then using R3 to mark AC and AD the same length). Then, if we connect C and D to B, we will be left with a situation like that shown in Figure 29. Marking AB with a new marker, and applying SAS to triangles CAB and DAB, we can mark CB and DB congruent with a new marker. Notice that we now have the same situation encountered in the proof of SSS and shown in Figure 28a, in which we have two different triangles with congruent sides on a single base. As before, we can show that this situation is impossible by reflecting the triangles over the base and then drawing two circles which would have to intersect in four places. This shows that SAS implies CA in **GS**, and so SAS and CA are equivalent in **GS**.

Similarly, we can define a system **GA** ("Geometry of Angles"), which contains all of the rules of **FG** except for CS, and a corresponding notion of **GA**-satisfaction in which $M \models D$ iff M's canonical unmarked diagram is equivalent to D's underlying unmarked diagram and if two di-angles in D are marked with the same marker, then the corresponding di-angles in M's canonical marked diagram are also marked with the same marker (so that here we have dropped the corresponding re-

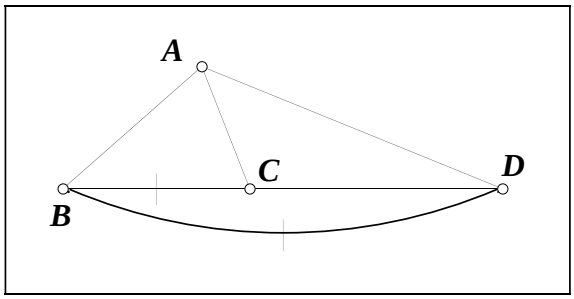


FIGURE 30 Deriving CS from SAS in **GA**.

quirement for dsegs). Again, all of the rules of **GA** are sound with respect to **GA**-satisfaction, but neither CA nor SAS are; this shows that neither CA nor SAS can be derived in **GA**. Furthermore, we can again show that SAS and CS are equivalent in **GA**. CS implies SAS in **GA** as before; again, this is shown by the proof that SAS is derivable in **FG**. So it suffices to show that CS is derivable from SAS in **GA**. To show this, let us be given a diagram in which one segment is contained in another with the same marking; call the first segment BC, and call the second BD. Next, pick (or construct) another point A that doesn't lie on the line BD. This gives the situation shown in Figure 30. Marking angle CBA and segment BA congruent to themselves with new markers, we can apply SAS to triangles CBA and DBA. This allows us to mark angle BAC congruent to angle BAD; but BAC is contained in BAD, and so we can eliminate this diagram by CA. This shows that CS is derivable from SAS in **GA**, and that SAS and CS are therefore equivalent in **GA**. This proof that CA and SAS together imply CS is identical to Hilbert's proof of the uniqueness of segment construction in Hilbert (1971).

Finally, we can look at a formal system that doesn't contain either CA or CS, but instead contains SAS. Let **BG** ("Basic Geometry") be the formal system containing all of the rules of **FG** except for CS and CA, and let **GSAS** ("Geometry of SAS") be **BG** with the added rule SAS. We have already shown that SAS and CS together imply CA in **BG**, and that SAS and CA together imply CS in **BG**, so this means that CA and CS are equivalent in **GSAS**. However, neither CS nor CA is derivable in **GSAS** without the other. To see this, define MM-satisfaction ("Meaningless Marker satisfaction") so that $M \models D$ iff M's canonical unmarked diagram is equivalent to D's underlying unmarked diagram. This allows any angle or segment to be marked

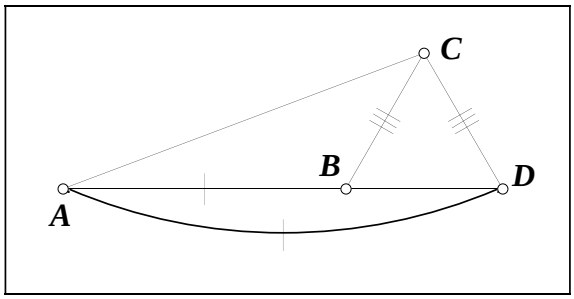


FIGURE 31 Deriving CS from SSS in **GA**.

the same as any other angle or segment, so that the markings have become meaningless. All of the rules of **BG** are sound with respect to MM-satisfaction, and so is SAS, because we can safely mark any dsegs or di-angles congruent without changing the models of a diagram; but neither CS nor CA are sound with respect to MM-satisfaction, since there are lots of diagrams satisfying the hypotheses of these rules which are still MM-satisfiable.

We have shown that the following interesting situation holds:

Theorem 2 *CS, CA, and SAS are independent of one another in* **BG***—that is, no one of them is provable from any other in* **BG***. However, any two of them are equivalent in the presence of the third, so that any one of them is provable from the other two.*

Notice that while SSS is provable from CS in **BG**, CS is not provable from SSS, because SSS is sound with respect to MM-satisfaction. So SSS is a weaker axiom than CS relative to **BG**. Relative to **GA**, however, the two axioms are equivalent: if we are given a diagram in which AB and AD are marked congruent and B lies on AD, as in Figure 31, we can construct an equilateral triangle on BD as in Euclid's first proposition. Calling the new vertex of this triangle C, we can connect C to A. If we mark AC with a new marker, we can apply SSS to triangles CBA and CDA. This allows us to conclude that angle ACB is congruent to angle ACD, which gives us the condition to apply CA and eliminate the diagram. So adding SSS and CA to **BG** gives us all of **FG**, while adding SSS and CS to **BG** just gives us **GS**. Adding SSS and SAS to **BG** gives us a system that is weaker than **FG**, because it is sound with respect to MM-satisfaction, but may be stronger than **GSAS**. (I conjecture but haven't proven that SSS isn't provable in **GSAS**.)

We could go on proving results like this for quite some time. For another example, the Isosceles Triangle Theorem (ITT), which says

that if two sides of a triangle ABC are congruent, then its corresponding angles are also congruent, is not provable in **BG**, but can be proven by applying either SSS or SAS to ABC and CBA, so it is provable in both **GSAS** and **GS**. ITT isn't provable in **BG** because it isn't valid with respect to **GA**-satisfaction (but all of the rules of **BG** are). On the other hand, SAS isn't provable in **BG** + ITT, since ITT is valid with respect to **GS**-satisfaction and SAS isn't.

A nice way to think about all of these results is in terms of the lattice of subtheories of **FG**.

Definition 11 A ***derivation theory*** is set $\mathcal{T}$ of pairs of diagram arrays with the property that if $(d_1, d_2) \in \mathcal{T}$ and $(d_2, d_3) \in \mathcal{T}$, then $(d_1, d_3) \in \mathcal{T}$.

Definition 12 The ***theory of diagrammatic formal system*** F, denoted Th(F), is the set of pairs (d_1, d_2) such that $d_1 \vdash d_2$ in F.

The set of subtheories of Th(**FG**), that is, the set of all subsets of Th(**FG**) which are themselves derivation theories, forms a lattice under the partial ordering given by the set inclusion relation. (A lattice is a partially ordered set in which any two elements a and b have a least upper bound, written "$(a \vee b)$" and called their ***join***, and a greatest lower bound, written "$(a \wedge b)$" and called their ***meet***.) All of the above results about the relationships between the various axioms CA, CS, SAS, SSS, and ITT can be restated as facts about this lattice. For example, the fact that CS (and therefore all of **FG**) can be derived from SAS and CA can be restated by saying that (Th(**GSAS**) $\vee$ Th(**GA**)) = Th(**FG**).

All of the above results can be put together to give the structure of part of this lattice, shown in Figure 32.

Thus, we see that there is an intricate and fascinating logical structure to diagrammatic proof systems in geometry, and it is one that is only apparent once we have formalized our proof methods.

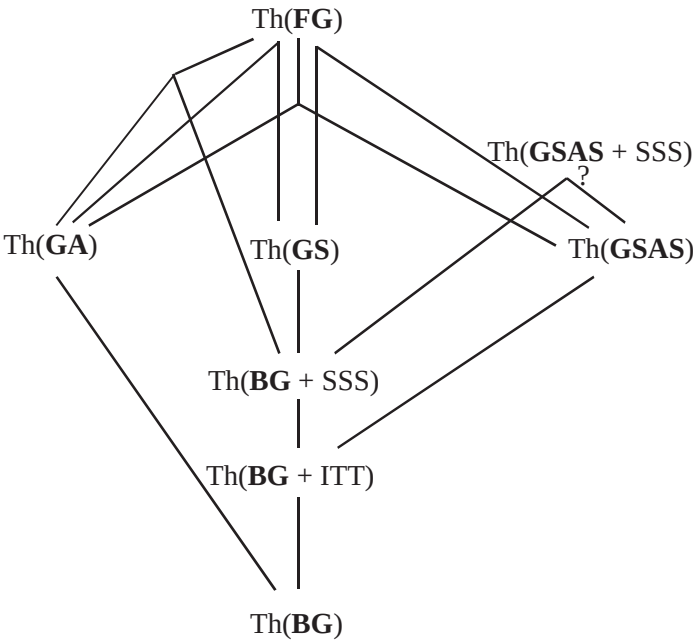


FIGURE 32 Part of the lattice of subtheories of Th(**FG**).

5

Conclusions

"There never has been, and till we see it we never shall believe that there can be, a system of geometry worthy of the name, which has any material departures (we do not speak of *corrections* or *extensions* or *developments*) from the plan laid down by Euclid." This was Augustus De Morgan's opinion in 1848, as cited and reaffirmed by Sir Thomas Heath in the preface to the 1901 first edition of his translation of the *Elements*, written two years after the appearance of Hilbert's *Foundations of Geometry*. However, since then, this view has virtually disappeared. As we have seen, Hilbert's geometry, which is very significantly different from Euclid's, has come to be seen as the true foundation of the subject. Since Hilbert's treatment of the subject meets modern standards of formality in a way that Euclid's does not, this may not be surprising.

What have been the practical consequences of the shift from Euclid to Hilbert as the generally accepted foundation for geometry? One might expect that once the foundations of a subject were finally, after two thousand years, put on a completely solid footing, the subject would flourish, and education about and understanding of the subject would improve. Sadly, however, nearly the opposite has been true. The study of Euclid has faded away as its claim to rigor has become suspect, but it has not been replaced by a study of Hilbert's geometry. In fact, knowledge of Hilbert's geometry is uncommon even among professional mathematicians.

Why is this? One reason is that Hilbert's axiom system, complete though it may be, is too abstract to be easily used by beginning students. Such students normally don't have a sophisticated enough understanding of the logic underlying mathematics to appreciate the point of using his axiom system. It should be said that a number of different curricula have been put forward to try to teach beginning undergrad-

uate students geometry based on (subsets of) Hilbert's axioms. I have taught classes using such materials, and my experience has been that, while beginning students can learn a lot about *logic* from such materials, they don't learn very much about *geometry*. Thus, while Hilbert's approach was groundbreaking from a logical point of view, it isn't well suited to the needs of beginning geometry students. In *Euclid and his Modern Rivals*, the protagonist, Minos, makes similar comments about Legendre's *Eléments de Géométrie*, much of which he says is, "though a model of elegance and perspicuity as a study for the advanced student, ... wholly unsuited to the requirements of a beginner."

So we see that study of Euclid's *Elements* faded because it was seen as unrigorous compared to Hilbert's geometry, but study of Hilbert's geometry didn't replace the study of the *Elements*, because his methods are too hard for beginners to use in learning geometry. This left somewhat of a vacuum in the study of beginning geometry, and while the subject is still studied, no clear consensus has emerged as to what it should include or how it should be taught, to say nothing of the question that was Dodgson's main concern, namely, what should be its precise logical development. In fact, we see that what has come to pass in the teaching of geometry is exactly what was prophesied by Dionysius Lardner, the editor of an 1828 edition of the *Elements*, who wrote in 1846 that, "Euclid once superseded, every teacher would esteem his own work the best, and every school would have its own class book. All that rigor and exactitude which have so long excited the admiration of men of science would be at an end. ... Every school would have a different standard; ... until, at length, GEOMETRY, in the ancient sense of the word, would be altogether frittered away or be only considered as a particular application of Arithmetic and Algebra."[20]

This seems even more prescient in light of the fact that the primary subject that has replaced geometry as most students' introduction to higher mathematics at the college level is Calculus, which can be viewed as precisely the study of Geometry as an application of Arithmetic and Algebra. Very few people, though, would argue that the study of calculus is a better place than Euclidean geometry to learn about the logic of mathematics, and what it means to do mathematics and to think mathematically. It is ironic but true, then, that the development of more rigorous logical methods in geometry contributed to a general decrease in the general study of and knowledge about both logic and geometry.

[20] As also quoted by Sir Thomas Heath in the preface to his 1901 edition of the *Elements*.

So what kind of implications does a system like the one developed in this book have for the actual practice of geometry as it is studied and taught? I think that it has several. First and foremost, it shows that there is, in fact, an underlying logic to Euclid's methods that can be made every bit as rigorous as other methods of presenting geometry. The idea that his methods are inherently informal should therefore be banished forever. (Not that this idea is likely to be eliminated any time soon.) Euclid's methods, as articulated in the *Elements*, are not completely formal, but requiring a work written two thousand years ago to completely live up to modern ideas of rigor is an impossible standard. In fact, the vast majority of actual mathematics, as it is practiced and taught, is not formal. So, really, Euclid's mathematics fits right in. It is just because it comes so close to being completely formal that people are tempted to apply such standards to this work. In any case, its blend of the formal with the informal makes the *Elements* a perfect introduction for beginners to mathematical standards and modes of argument, as much so now as it has been for two millennia.

However, restoring Euclid's reputation is decidedly not the only reason for looking at a formal system like ours. There are, in fact, many other reasons that looking a system like this is fruitful.

First of all, it has great value for anyone who wants to understand the logic underlying informal geometric proofs. For example, it tells us the proper rules by which geometric diagrams can be carefully and correctly used. Even if we may choose to ignore some of the rules in practice (after all, it's hard to do the required case analysis by hand), at least we will have a good idea what we would have to check in order to make sure that a proof is valid. This is, after all, the same situation that holds in the rest of mathematics, in which people normally give informal proofs, checking enough details to convince themselves that the proof could be made formal.

Next, a formalism like this shows once and for all that geometric proofs that use diagrams are in no way inherently less rigorous than sentential proofs. The two different styles of proofs certainly have different strengths and weaknesses, but neither can lay a greater claim to inherent rigorousness. It is possible to prove fallacies using diagrams, but only if they are used incorrectly, just as it is possible to prove that $0 = 1$ by using algebra incorrectly. Thus, the twentieth century bias against the use of diagrams in geometry can be seen for what it is—a bias, which hopefully will slowly dissipate. Then, instead of viewing the use of a diagram in a proof as a mark of "human frailty," people can view diagrams as useful tools to be understood and used wisely. After all, people are going to use diagrams in their proofs in either case.

It should be noted that this is certainly not the first work to show that diagrams can can be used rigorously in mathematics, and yet the bias against diagrams is still strong. However, I think that it is less strong than it once was. Furthermore, as we have previously discussed, there are perfectly good reasons why a skeptic might doubt that the use of diagrams in Euclidean geometry could be made rigorous even if that skeptic was familiar with other work formalizing the use of diagrams. We have seen that the diagrams used in geometry are more complex than other kinds of diagrams that have been formalized, and therefore harder to abstract correctly. Furthermore, many of the other areas in which the use diagrams has been formalized, such as Venn Diagrams as formalized by Shin, or Hyperproof's blocks world, are used for teaching the use of logic in mathematics, but they are not commonly used in ordinary day to day mathematics. Thus, the particular formal system developed in this book is useful because it shows that diagrams can be used rigorously not only in carefully constructed teaching environments, but also in the somewhat messier world of day to day mathematics.

The next reason that a formal system like this is useful is that it helps to explain some the ways in which people have traditionally used diagrams. For example, the observation that the use of lemmas can lead to an exponential decrease in the number of cases that need to be considered helps to explain why geometry has such a long history of basing proofs on collections of previously proven facts. It also provides an alternative explanation for the fact that Euclid used superposition to prove SAS and then used SAS to prove other results rather than continuing to use superposition. Many commentators have asserted that this shows that Euclid viewed superposition as being a suspect method of proof. While it is possible that this was the case, the formal system **FG** shows that proofs that use superposition can be made as rigorous as other proofs, and that there are still good reasons for preferring the use of SAS as a lemma to the direct use of superposition in general. **FG** also sheds light on the old dispute over how many different cases need to be considered in proving a theorem. Euclid's normal practice was to give the proof for a single case only, but many later commentators have pointed out that there can be other cases that need to be considered, often with corresponding changes in the proof. It has not been previously clear exactly how many cases needed to be considered, which has contributed to the idea that the use of diagrams is inherently informal. (In fact, there has sometimes been disagreement over this: see for example Thomas Heath's commentary on Euclid's Proposition 2,[21]

[21] (Euclid, 1956, pp. 145–146)

in which he discusses Proclus' case analysis, poking fun at his "anxiety to subdivide [cases]."[22]) The semantics of diagrams that we have given here makes it clear that two cases are distinct and may require different proofs if they are topologically different. Case analysis aside, it is striking how similar proofs in **FG** can be to those given by Euclid. In fact, many of Euclid's proofs that have been often criticized for making unstated assumptions, such as the proof of his first proposition, turn out to look exactly the same in **FG**, because the assumptions are taken care of by the underlying diagrammatic machinery. Thus, **FG** shows that some of the aspects of Euclid's proofs that have been viewed as flaws can be viewed as correct uses of a diagrammatic method that was not fully explained.

Finally, a system like this is enormously helpful in proving metamathematical results about geometry. In fact, questions like "How hard is it to determine if a given diagram is satisfiable?," "Is CS a stronger axiom than SSS?," or "Are there true facts about 2-dimensional Euclidean geometry that cannot be proven by standard Euclidean methods?" can't even be articulated clearly until we have a formalization like this. (Another way of stating the third of these questions is "Is $\mathbf{FG}'$ complete?," and it is an open question.) The development of **FG** and $\mathbf{FG}'$ has allowed these kinds of questions to be meaningfully articulated; some have been answered here, but many more remain.

[22]See also Manders (1995) for a discussion of the historical role of commentators in proposing additional cases that need to be considered.

Appendix A

Euclid's Postulates

For reference, Euclid's Postulates and Common Notions, along with several of his definitions from Book I of *The Elements*, are given in Tables 10, 11, and 12, as translated in Euclid (1956).

TABLE 10 Some of Euclid's definitions from Book I of *The Elements*.

Definitions
1. A point is that which has no part.
2. A line is breadthless length.
10. When a straight line set up on a straight line makes the adjacent angles equal to one another, each of the equal angles is ***right***, and the straight line standing on the other is called a ***perpendicular*** to that on which it stands.
15. A ***circle*** is a plane figure contained by one line such that the straight lines falling upon it from one point among those lying within the figure are equal to one another;
16. And the point is called the ***center*** of the circle.

TABLE 11 Euclid's Postulates from *The Elements*.

Postulates

Let the following be postulated:

1. To draw a straight line from any point to any point.
2. To produce a finite straight line continuously in a straight line.
3. To describe a circle with any center and distance.
4. That all right angles are equal to one another.
5. That, if a straight line falling on two straight lines make the interior angles on the same side less than two right angles, the two straight lines, if produced indefinitely, meet on that side on which are the angles less than the two right angles.

TABLE 12 Euclid's Common Notions from *The Elements*.

Common Notions

1. Things which are equal to the same thing are also equal to one another.
2. If equals be added to equals, the wholes are equal.
3. If equals be subtracted from equals, the remainders are equal.
4. Things which coincide with one another are equal to one another.
5. The whole is greater than the part.

Appendix B

Hilbert's Axioms

In this section we reproduce Hilbert's axioms for geometry, as given in Hilbert (1971). Hilbert divides his axioms into five different groups.

TABLE 13 Hilbert's Axioms of Incidence

Axioms of Incidence
I, 1. For every two points A, B there exists a line a that contains each of the points A, B.
I, 2. For every two points A, B there exists no more than one line that contains each of the points A, B.
I, 3. There exist at least two points on a line. There exist at least three points that do not lie on a line.
I, 4. For any three points A, B, C that do not lie on the same line there exists a plane α that contains each of the points A, B, C. For every plane there exists a point which it contains.
I, 5. For any three points A, B, C that do not lie on one and the same line there exists no more than one plane that contains each of the three points A, B, C.
I, 6. If two points A, B of a line a lie in a plane α then every point of a lies in the plane α.
I, 7. If two planes α, β have a point A in common then they have at least one more point B in common.
I, 8. There exist at least four points which do not lie in a plane.

TABLE 14 Hilbert's Axioms of Order

Axioms of Order	
II, 1.	If a point B lies between a point A and a point C then the points A, B, C are three distinct points of a line, and B then also lies between C and A.
II, 2.	For two points A and C, there always exists at least one point B on the line AC such that C lies between A and B.
II, 3.	Of any three points on a line there exists no more than one that lies between the other two.
II, 4.	Let A, B, C be three points that do not lie on a line and let a be a line in the plane ABC which does not meet any of the points A, B, C. If the line a passes through a point of the segment AB, it also passes through a point of the segment AC, or through a point of the segment BC.

TABLE 15 Hilbert's Axioms of Congruence

Axioms of Congruence	
III, 1.	If A, B are two points on a line a, and A' is a point on the same or on another line a' then it is always possible to find a point B' on a given side of the line a' through A' such that the segment AB is congruent or equal to the segment $A'B'$. In symbols, $AB \cong A'B'$.
III, 2.	If a segment $A'B'$ and a segment $A''B''$ are congruent to the same segment AB, then the segment $A'B'$ is also congruent to the segment $A''B''$, or briefly, if two segments are congruent to a third one they are congruent to each other.
III, 3.	On the line a let AB and BC be two segments which except for B have no point in common. Furthermore, on the same or on another line a' let $A'B'$ and $B'C'$ be two segments which except for B' also have no point in common. In that case, if $AB \cong A'B'$ and $BC \cong B'C'$, then $AC \cong A'C'$.
III, 4.	Let $\angle(h,k)$ be an angle in a plane α and a' a line in a plane α‘ and let a definite side of a' in α' be given. Let h' be a ray on the line a' that emanates from the point O'. Then there exists in the plane α' one and only one ray k' such that the angle $\angle(h,k)$ is congruent or equal to the angle $\angle(h',k')$ and at the same time all interior points of the angle $\angle(h',k')$ lie on the given side of a'. Symbolically, $\angle(h,k) \cong \angle(h',k')$. Every angle is congruent to itself, *i. e.*, $\angle(h,k) \cong \angle(h,k)$.
III, 5.	If for two triangles ABC and $A'B'C'$ the congruences $AB \cong a'B'$, $AC \cong A'C'$, $\angle BAC \cong \angle B'A'C'$ hold, then the congruence $\angle ABC \cong \angle A'B'C'$ is also satisfied.

TABLE 16 Hilbert's Axiom of Parallels

Axiom of Parallels
IV. Let a be any line and A a point not on it. Then there is at most one line in the plane, determined by a and A, that passes through A and does not intersect a.

TABLE 17 Hilbert's Axioms of Completeness

Axioms of Completeness
V, 1. **(Archimedes' Axiom)** If AB and CD are any segments then there exists a number n such that n segments CD constructed contiguously from A, along the ray from A through B, will pass beyond the point B.
V, 2. **(Axiom of line completeness)** An extension of a set of points on a line with its order and congruence relations that would preserve the relations existing among the original elements as well as the fundamental properties of line order and congruence that follows from Axioms I–III, and from V, 1 is impossible.

Appendix C

Isabel Luengo's DS1

This appendix contains a summary of Isabel Luengo's formal system **DS1** for geometry and an explanation of why it is unsound. Her description of this system can be found in Chapter VII ("A Diagrammatic Subsystem of Hilbert's Geometry") of the book *Logical Reasoning with Diagrams*,[23] which contains the same material (slightly abbreviated) as Chapters 2 and 3 of her thesis.[24]

First, her syntactic objects. These are geometric objects in the plane. She recognizes four kinds of primitive syntactic objects: Boxes, Points*, Lines*, and Indicators. A Box is a dashed rectangle, a point* is a dot, a line* is a (genuinely) straight line in the plane, and an indicator is a collection of slash marks, possibly sitting on an arc. (The indicators will be used to mark segments as having equal lengths.) She recognizes four relations that can hold between diagrammatic objects: In, which tells you if a given object lies entirely inside a box; On*, which tells you if a point* intersects a line*; Indicates, which tells you if an indicator indicates a pair of points*; and Between* (this is the one that will turn out to be important in the subsequent discussion), which is defined as follows: "Point* A is between* B and C if and only if there is a line* that goes through* A, B, and C, and A is between B and C on that line*." Note that A cannot be between* B and C unless there is a line* drawn through A, B, and C.

Next, she defines a diagram to be any finite combination of primitive diagrammatic objects, and a well-formed diagram to be one which contains a single box, such that all other primitive diagrammatic objects are In the box, such that given any two distinct points* there is at most one line* through them, and such that every indicator indicates a segment (where segment is a derived term meaning two points* on a

[23] Barwise and Allwein (1996)
[24] Luengo (1995)

line* l and including the part of the line* between them; an indicator indicates a segment iff it indicates the pair of points). She then defines two diagrams to be copies of one another iff there is a bijection between them preserving the four relations In, On*, Between*, and Indicates. (She actually gives a slightly more complicated equivalent definition.) She shows that this is an equivalence relation, and says that from this point forward, D will mean the equivalence class of all diagrams that are copies of D. Note that this notion of equivalence doesn't contain any topological information about how the diagram lies in the plane. In her dissertation, Luengo also discusses a formal system **DS2** that includes the relation of a ray being in the interior of an angle, and several extensions of this formal system that include the relation of points lying on the same side of a given line. These formal systems do contain some topological information. We won't discuss them further here; but they are based on **DS1**, and the examples given below that show that **DS1** is unsound also apply to these extensions.

Now, the semantics of **DS1**. A function $I : D \to E$ is an ***interpretation function*** for D iff it is a total function from the points* and lines* of D to the points and lines of E, where E is a Euclidean plane (which is defined to be anything satisfying Hilbert's axioms), such that I takes points* in D to points in E, lines* in D to lines in E, such that point* A is on* line* L iff $I(L)$ goes through point $I(A)$, and (here's the key part) such that if A, B, and C are points* then A is between* B and C iff $I(A)$ is between $I(B)$ and $I(C)$. This definition is where the trouble first arises: note that because the definition of between* requires the points* to lie on a common line*, if D is a diagram containing three points* that don't lie on any common line*, then I isn't an interpretation function for D if $I(A)$, $I(B)$, and $I(C)$ are collinear. If $I(A)$, $I(B)$, and $I(C)$ are collinear with $I(A)$ between $I(B)$ and $I(C)$ then we would have to have A between* B and C; but it can't be, because there is no line* l that A, B, and C all lie on*. (Intuitively, it would make more sense just to require that if A is between* B and C, then $I(A)$ is between $I(B)$ and $I(C)$. If I satisfies this weaker condition, call it a ***weak interpretation function***.)

If I is an interpretation function for D, then it is a ***pre-model*** if 1) if two segments S_1 and S_2 are marked equal in D, then $I(S_1)$ and $I(S_2)$ are equal in length; and 2) if A and B are on the same side of line* l with respect to point* C, then $I(A)$ and $I(B)$ are on the same side of $I(l)$ with respect to $I(C)$. A and B are defined to be on the same side of L with respect to C iff A, B, and C are all on* L and C isn't between* A and B. So (2) is equivalent to saying that if A, B, and C all lie on* L and C isn't between* A and B, then $I(A)$, $I(B)$, and $I(C)$ all lie on

$I(L)$, and $I(C)$ isn't between $I(A)$ and $I(B)$. This is already part of the definition of being an interpretation function, but it wouldn't be under the weaker definition of an interpretation function. Note that if I is a weak interpretation function satisfying these two conditions (call such a function a ***weak pre-model***), then if $I(C)$ is between $I(A)$ and $I(B)$ and A, B, and C all lie on a common line* L, then C must lie between* A and B, by condition (2).

Luengo then defines three "deduction principles" (later called construction rules). The first of these is rule P1 (line* introduction): given a diagram D with points* A and B not on a single line* of D, one can deduce a diagram E that is "just like" D except that it has a new line* through A and B. There are a number of different ways that this rule can be construed. Because they are somewhat involved, I'm going to defer discussing them until we need to use this rule later on. The other two rules are rule P2 (point* introduction): given D containing a line L, deduce any diagram E identical to D, but with a new point* that is on* L and not on any other line*; and rule P3 (existence of segments): this is a more complicated rule that says that given a marked segment length M, a point* A on a line* L, and a given side D of A on L, you can add a new point* B on* L, on the given side of A, and mark the segment from A to B with marker M. If there are other points on the given side of A, you get a disjunction of diagrams showing the different ways B could lie on L with respect to the existing points.

A premodel M of D is defined to be a ***model*** of D if for any diagram E (or disjunctive set of diagrams S) obtainable from D via one of the three deduction principles, M can be extended to a premodel of E (or S).

We now reach a central proposition of her paper, Proposition 3.8: A pre-model M is a model iff it is 1-1. Both directions of this proposition are false.

The if direction is false because we can find diagrams D and D^* such that D has a 1-1 premodel, D^* is constructible from D, and D^* doesn't have any premodels, because it contains three points* that don't lie on a common line*, but would have to be collinear in any possible premodel. As noted above, if A, B and C are points* in D^* that don't lie on a common line*, then if $I(A)$, $I(B)$, and $I(C)$ are collinear, then I can't be an interpretation function for D^*. So in this case D^* couldn't have any premodels.

Figure 33 shows an example of how this can happen. Diagram D has a 1-1 premodel that is the one that you would expect it to have, and diagram D^* is obtainable from diagram D by rule P3. But in any possible premodel of D^*, p, q, and r would have to be collinear–they

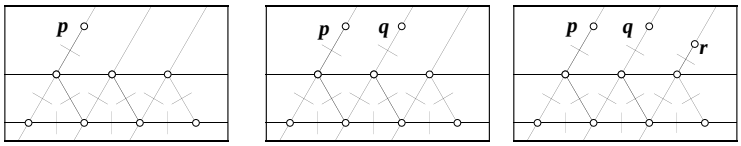


FIGURE 33 A counterexample to the soundness of **DS1**.

each sit the same distance from L along lines that make a 60 degree angle with L. So as discussed above, D^* can't have any premodels in her system, since there is no line through p, q, and r. But D^* is derivable from D, which has a 1-1 premodel; so the if direction of the proposition is false. Notice that more or less the same example shows that her system is in fact unsound. D doesn't have a model by her definition (because D^* doesn't have a premodel), but it is constructible from diagram D'', which does have a model. Since D'' has a model and D doesn't, but D is constructible from D'', the system is unsound. Notice that this counterexample doesn't rely on our interpretation of rule P1; so far, we have only used rule P3.

The only if direction of Proposition 3.8 can also be seen to be false as follows: let E be a diagram containing two points* p_1 and p_2 (and nothing else). Let M be a premodel of E such that $M(p_1) = M(p_2)$, so M isn't 1-1. The only deduction principle that applies is Line* introduction. So let E' be a diagram obtained from E by adding a line* L through p_1 and p_2. Let $N(p_1) = N(p_2) = M(p_1) = M(p_2)$, and let $N(L)$ be some line going through $N(p_1)$. N is a premodel of E'. So M was a non-1-1 model of E. The proof that Luengo gives here actually shows something different: that if M is a non-1-1 premodel of E, then there is a diagram E'' that can be obtained from E by applying a sequence of deduction principles, such that M can't be extended to a pre-model of E''. In any case, this direction of the proposition isn't as important, as we can simply require all interpretation functions to be 1-1.

The above example suggests that her system might be made sound by only requiring premodels to have weak interpretation functions. Under the weaker definition, diagram D^* above would have a premodel. Unfortunately, there would still be a problem. Luengo proves in her Proposition 6.5 that with this change her system is not sound. The problem is with rule P1. Recall rule (P1): given a diagram D with points* A and B not on a single line* of D, one can deduce a diagram E that is "just like" D except that it has a new line* through A and

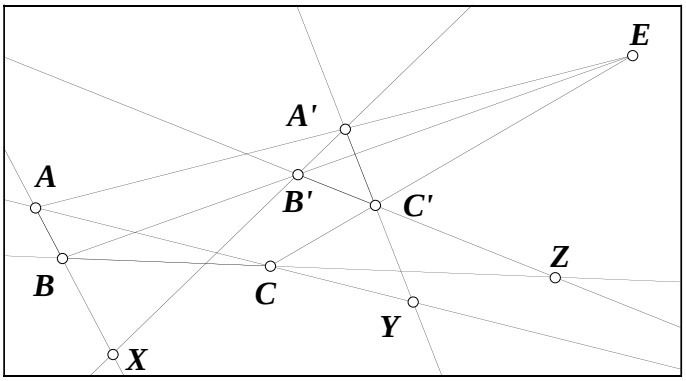


FIGURE 34 Desargues' theorem.

B.

As previously remarked, there are a number of different ways that this rule can be construed. Here are four possibilities:

(Version 1) E is a diagram containing a new line* l such that A and B are on* l and no other point* C in E is on* l, and removing l from E leaves a diagram equivalent to D. This way of construing the rule takes the words "just like" to be understood to be referring to the On* relation. An immediate problem here is that it isn't obvious whether or not we can always find such a diagram E. In fact, we can't. A counterexample is given by Desargues' Theorem, the proof of which was discussed in Section 4.2. Let D be a diagram obtained as follows: Draw a triangle ABC. Draw a point E somewhere not on the triangle. Draw in the lines AE, BE, and CE. Pick points A' on AE, B' on BE, and C' on CE. Connect A', B', and C' to obtain a new triangle $A'B'C'$. Extend the six line segments that make up the sides of the two triangles to lines. Mark point X at the intersection of lines AB and $A'B'$, point Y at the intersection of lines AC and $A'C'$, and point Z at the intersection of lines BA and $B'A'$. What you have now is diagram D, shown in Figure 34. Desargues' theorem says that in D (and in any equivalent diagram) points X, Y, and Z must be collinear. If we want to apply rule P1 to points X and Y, there is no way to draw a straight line through X and Y that doesn't also go through Z. So if we want to construe the rule this way, we'll have to accept that it won't be possible to apply the rule to all diagrams that have two points* not on a common line*.

(Version 2) The rule could also mean that E is the diagram in which A and B are really connected by a new straight line, and any other

point* C is on* the new line iff it happened to intersect the real straight line between A and B in D. We get this version of the rule if we take the words "just like" to refer literally to D as a geometric object itself. This version of the rule suffers from a different problem: it can be applied to any diagram in which there were two points* A and B that didn't lie on a common line*, but it isn't well defined on equivalence classes of copies of diagrams. For example, the point* C might lie directly between A and B in D, but not in some other diagram D' that was equivalent to D. To fix this problem, we can modify the rule as follows:

(Version 3) E is obtainable from D by P1 as long as there is some diagram D' that is equivalent to D such that E can be obtained from D' by connecting A and B by a straight line (and, as in the previous version, any other point* C lies on* the new line iff C really lies between A and B in D'). I think that this is the best way to interpret the rule. It doesn't suffer from either of the above problems with versions one and two. In fact, if E is obtainable from D by either Version 1 or Version 2 of the rule, it will also be obtainable by Version 3. Version 3 also has the property that, unlike the previous versions, there are multiple diagrams that can be obtained from a given diagram by applying rule P1 to two given points. This seems to be what Luengo intends. In her thesis (p. 23), she writes that this rule states that "any extension of the diagram that meets a certain condition is obtainable from the diagram."

There is one other version of the rule that might be suggested:

(Version 4) E is obtainable from D iff for every diagram D' that is equivalent to D, adding the straight line that runs through points* A and B gives a diagram E' that is equivalent to E. This is equivalent to saying that E is obtainable from D by Version 3 of the rule, and is the only diagram obtainable from D by Version 3. This version of the rule shares Version 1's property that there sometimes isn't any diagram that can can be derived from D by applying the rule to two given points* that don't lie on any common line*. (This will happen any time that there are two or more diagrams derivable by Version 3 of the rule.) More importantly, though, in order to apply this rule, one has to check that for every diagram D' that is a copy of D, the result of drawing in the straight line through the given points* is a copy of E. There isn't an obvious general method for doing this; we can't check directly, because there are usually an infinite number of different copies of D. And in the cases where we can in fact show that this property holds for every such copy D', the proof may involve a great deal of prior geometric knowledge, as in the example given for Version 1, in which we had to apply Desargues' theorem. Obviously, it is highly undesirable to have to use complicated geometric facts to determine if

one diagram follows from another syntactically in a formal system for doing geometry. It seems to me that this problem makes this version of the rule unworkable in practice. (This version also seems inconsistent with Luengo's own use of this rule, for example in her Proposition 6.5.)

Thus, the version of P1 that seems to have the fewest problems and seems most consistent with Luengo's intent is Version 3. Unfortunately, under the weak definition of premodel, this rule is unsound. In fact, Versions 1–3 of P1 are all unsound. To see this, consider diagram D^* from above, and assume that point r has been drawn a different distance above the triangles than p and q were drawn. Then by applying any of the first three versions of rule P1, to points p and q in D^*, we can obtain a diagram D''' with a line* L running through p and q, but not r; but D''' can't have any premodels (even under the weaker definition), since in any premodel, $I(L)$ would have to run through $I(r)$ if it ran through $I(p)$ and $I(q)$. But D''' was obtained from D^*, which does have (weak) 1-1 premodels. Notice that with the original definition of premodel you don't run into the second problem, because your model never contains extra points that might show up on the new line that you're constructing.

So, under the weaker definition of premodel, Versions 1–3 of rule P1 are unsound. Version 4, on the other hand, is sound: any model M of a diagram D is itself a copy of D once we add in a box and indicators, so if E is obtained by applying Version 4 of P1 to points* A and B in D, and N is the extension of M in which the line L through $M(A)$ and $M(B)$ has been added, then N satisfies E by the definition of Version 4 of rule P1; but this version of the rule is unworkable in practice, as discussed above.

There is actually one other possible modified version of rule P1 that is sound and slightly better than Version 4. I'll call this modified version P1′: given a diagram D with points* A and B not on a single line* of D, one can deduce the disjunctive set of diagrams $S = \{E|$ there exists a copy D' of D such that E is obtained from D' by adding the straight line through A and $B\}$. Rule P1′ is sound, again because any model M of D will give you a copy D' of D. Unlike Version 4 of rule P1, given any diagram D with points* A and B not on* a common line*, there always exists a disjunctive set of diagrams S such that S follows from D by rule P1′ applied to A and B. Intuitively, this modified rule makes sense. If we connect two points by a line, we have no way of knowing in advance which other points the new line will intersect, but we know that there there will be some way of connecting the two points by a straight line. However, P1′ is still unworkable in practice, because in general, we have no way of checking that we've found all of the diagrams

in S.

The only evident way that we can get around this difficulty is to relax the requirement that lines* be actually straight. If we don't require the lines* to be straight, then we can combinatorially compute the possible ways that the new line* could intersect the old points*, and using non-straight lines we can definitely realize all these possibilities. This is more or less the way that **FG** handles this issue. Using straight lines introduces too much geometric information into the diagram.

Appendix D

A CDEG transcript

The following is a transcript of a **CDEG** session, for readers who are interested in more details.

CDEG is driven by a text interface, and also has the ability to open windows with pictures displaying the diagram being worked on.

First we start **CDEG** and ask it what commands are available:

```
Welcome to CDEG 1.0!
(type h for help)
CDEG(1/1)% h
Options are:
<s>ave, <l>oad, se<t> pd, <v>iew current pd,
<a>dd dot to segment, add dot to <r>egion,
con<n>ect dots, <d>raw circle, <p>rint diagram,
<e>rase diagram, <m>ark radii, <c>ombine markers,
e<x>tend segment, add mar<k>ers, get <h>elp, <q>uit
```

The prompt here (`CDEG(1/1)%`) tells us that we are currently working with the first primitive diagram in a diagram array that contains 1 primitive diagram. Since we have just started the program, this is the empty primitive diagram. We can view it by typing "v":

```
CDEG(1/1)% v
```

This causes **CDEG** to open a window showing the diagram. This diagram is shown in Figure 35.[25] It contains a single region bounded by the frame; **CDEG** has assigned this region the number 4. **CDEG** as-

[25] As previously mentioned in Section 3.5, all of the **CDEG** diagrams included in this book have been reproduced as they were output by the **CDEG** program, but with two minor modifications: (1) the colors that **CDEG** uses to identify different lines and circles have been changed to different shades of gray; and (2), the locations of the numbers labeling the different segments have been altered in order to make the numbers more legible, since **CDEG** often places them so that they are obscured by the dots at the ends of the segments.

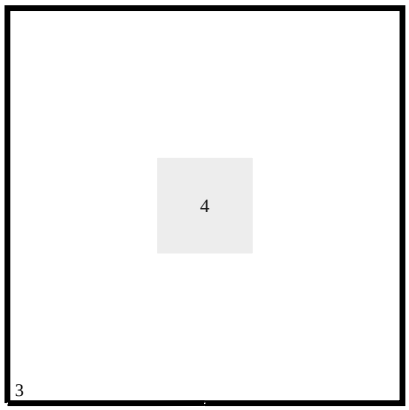


FIGURE 35 The empty primitive diagram as drawn by **CDEG**.

signs each object in a diagram a unique number by which it can be identified. Next, we use the "r" command ("`add dot to <r>egion`") to add two new dots to this region:

```
CDEG(1/1)% r
Enter region number: 4
CDEG(1/1)% r
Enter region number: 4
```

Now, let's look at the resulting diagram. We could use the `<v>iew` command to look at the resulting diagram, but we are going to instead use the `<p>rint diagram` command to print a text representation of the diagram. As discussed in section 3.5, this is essentially a version of the diagram's corresponding graph structure.

```
CDEG(1/1)% p
Diagram #1:
dot13 is surrounded by: region4
dot12 is surrounded by: region4
frame3 ends at loop in regions region4 and outerregion

region4 has boundry: frame3
        and contents:
Component #1: dot13
Component #2: dot12
```

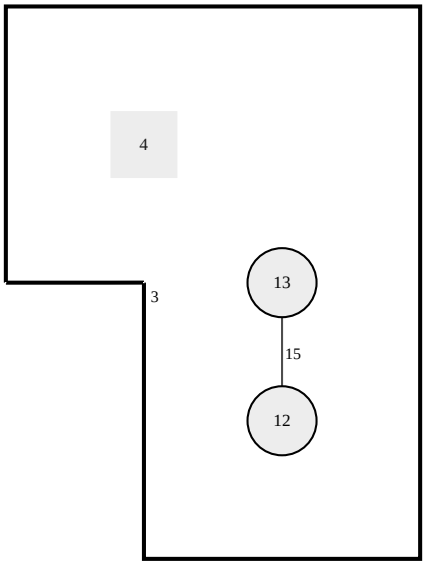


FIGURE 36 A **CDEG** diagram showing a single line segment.

So we see that the two new dots are numbered 12 and 13. We can connect them using the `con<n>ect dots` command.

```
CDEG(1/1)% n
Enter first dot's number: 12
Enter second dot's number: 13
CDEG(1/1)% v
```

The resulting diagram is shown in Figure 36. We will `<s>ave` this diagram so that we can come back to it, and then `<d>raw a circle` centered at dot 12 and going through dot 13.

```
CDEG(1/1)% s
Enter file name: seg.cd
CDEG(1/1)% d
Enter center dot's number: 12
Enter radius dot's number: 13
CDEG(1/1)% v
```

The resulting diagram is shown in Figure 37. The resulting dcircle is not at all circular, but all we care about here is the topology of the diagram. Next, we want to draw another circle centered at dot 13 and

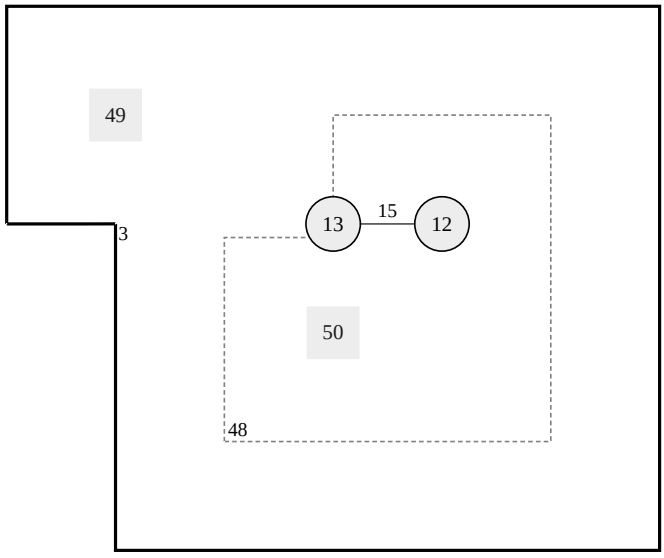


FIGURE 37 A **CDEG** diagram showing the second step in the proof of Euclid's First Proposition.

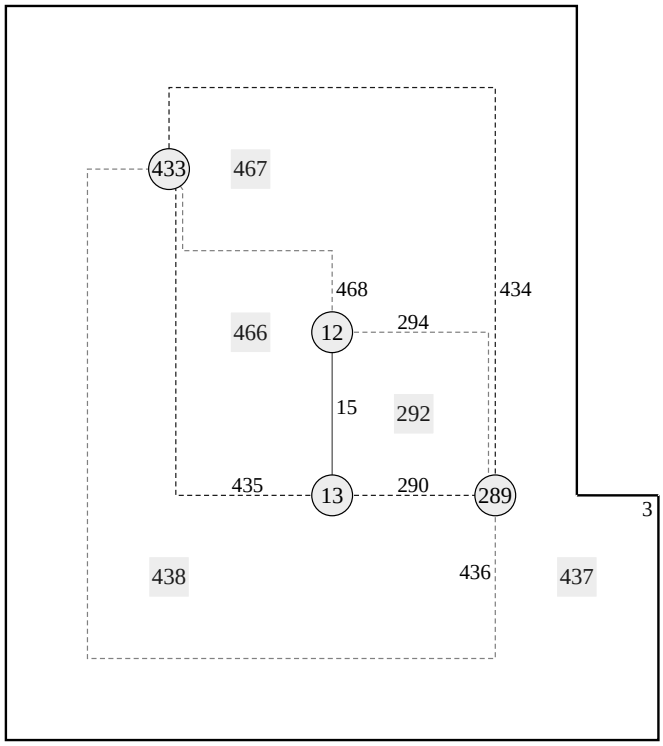


FIGURE 38 A **CDEG** diagram showing the third step in the proof of Euclid's First Proposition.

going through dot 12.

```
CDEG(1/1)% d
Enter center dot's number: 13
Enter radius dot's number: 12
CDEG(1/1)% v
```

This diagram is shown in Figure 38. Although it can't be seen as easily in the black and white printed version, the segments in each diagram are colored differently depending on which line or circle they are part of.

Next, we will form a triangle by connecting the endpoints of the segment to one of the points, dot number 289, on the intersection of the two circles.

```
CDEG(1/1)% n
Enter first dot's number: 12
Enter second dot's number: 289
CDEG(1/1)% n
Enter first dot's number: 13
Enter second dot's number: 289
CDEG(1/1)% v
```

The resulting diagram is shown in Figure 39. If we print out this much more complicated diagram, it looks like this:

```
CDEG(1/1)% p
Diagram #1:
dot433 is surrounded by: dottedseg434 region437
     dottedseg436 region438 dottedseg435 region466
     dottedseg468 region467
dot289 is surrounded by: dottedseg290 region438
     dottedseg436 region437 dottedseg434 region467
     dottedseg294 region1461 solid1462 region1762
     solid1763 region1761
dot13 is surrounded by: region466 dottedseg435 region438
     dottedseg290 region1761 solid1763 region1762 solid15
dot12 is surrounded by: region1762 solid1462 region1461
     dottedseg294 region467 dottedseg468 region466 solid15
solid1763 ends at dots dot13 and dot289
solid1462 ends at dots dot12 and dot289
dottedseg468 ends at dots dot433 and dot12
dottedseg436 ends at dots dot289 and dot433
dottedseg434 ends at dots dot289 and dot433
dottedseg435 ends at dots dot433 and dot13
dottedseg294 ends at dots dot12 and dot289
dottedseg290 ends at dots dot13 and dot289
solid15 ends at dots dot12 and dot13
frame3 ends at loop in regions region437 and outerregion
dline1463 is made up of dot289 solid1763 dot13
dline484 is made up of dot289 solid1462 dot12
dline14 is made up of dot13 solid15 dot12
circle87 has center dot13 and boundry dottedseg468 dot433
     dottedseg436 dot289 dottedseg294 dot12
circle23 has center dot12 and boundry dottedseg290 dot289
     dottedseg434 dot433 dottedseg435 dot13

region1761 has boundry: solid1763 dot13 dottedseg290
```

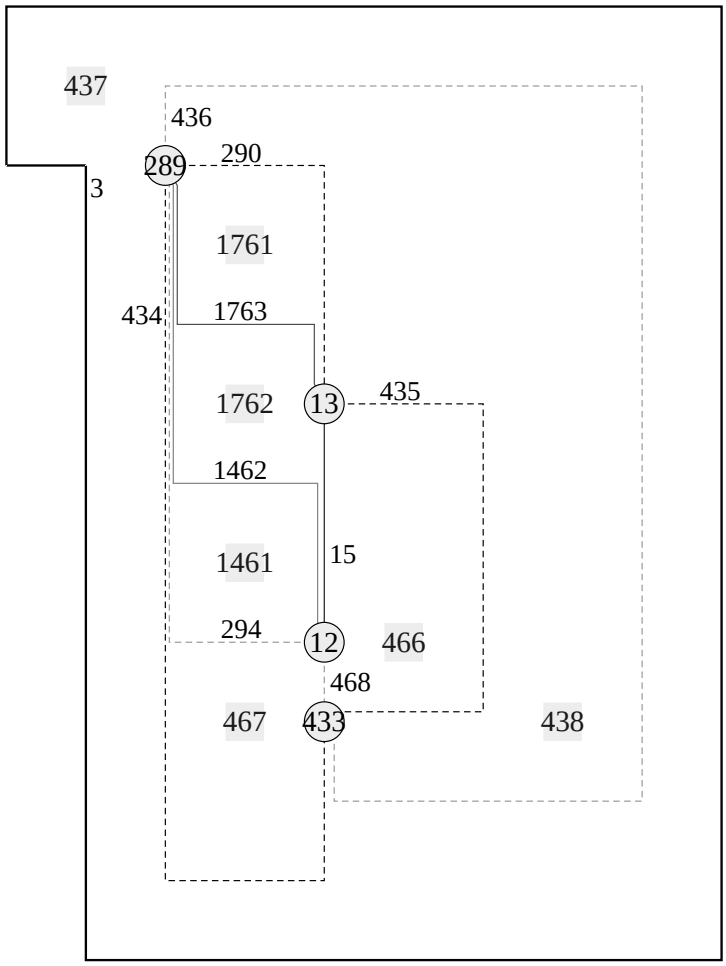


FIGURE 39 A **CDEG** diagram showing the triangle obtained in the proof of Euclid's First Proposition.

```
     dot289
        and contents:

region1762 has boundry: solid1763 dot289 solid1462 dot12
     solid15 dot13
        and contents:

region1461 has boundry: solid1462 dot289 dottedseg294
     dot12
        and contents:

region466 has boundry: dottedseg468 dot433 dottedseg435
     dot13 solid15 dot12
        and contents:

region467 has boundry: dottedseg468 dot12 dottedseg294
     dot289 dottedseg434 dot433
        and contents:

region438 has boundry: dottedseg436 dot289 dottedseg290
     dot13 dottedseg435 dot433
        and contents:

region437 has boundry: frame3
        and contents:
Component #1: dottedseg436 dot433 dottedseg434 dot289
```

Note that each dot lists the regions and segments that surround it in clockwise order; each segment lists its endpoints; each line and circle lists the dots and segments that make it up; and each region lists the segments and dots that are found around its boundary in clockwise order, and the segments and dots that make up the boundary of any connected components that are found inside the region.

Next, we want to mark segment 1763 congruent to segment 15. We can do this using the `<m>ark radii` command. This command lets us mark congruent two radii of the same circle; in order to use it, we must identify the circle that the radii are part of by identifying one of the segments that make it up. The radii are given as a list of the segments that make them up (since one radius may be made up of several diagrammatic pieces). **CDEG** checks to make sure that the given segments are in fact radii of the specified circle before it marks them; if they aren't, it returns an error message.

```
CDEG(1/1)% m
Enter the number of one seg on the circle: 468
Enter first radius dseg number:
Enter next seg index, or 0 to quit:1763
Enter next seg index, or 0 to quit:0
Enter second radius dseg number:
Enter next seg index, or 0 to quit:15
Enter next seg index, or 0 to quit:0
```

Similarly, we can mark segment 1462 congruent to segment 15 because they are both radii of the other circle.

```
CDEG(1/1)% m
Enter the number of one seg on the circle: 435
Enter first radius dseg number:
Enter next seg index, or 0 to quit:15
Enter next seg index, or 0 to quit:0
Enter second radius dseg number:
Enter next seg index, or 0 to quit:1462
Enter next seg index, or 0 to quit:0
CDEG(1/1)% v
marker2480 marks DSeg(solid15) DSeg(solid1462)
marker2479 marks DSeg(solid1763) DSeg(solid15)
```

The diagram that is displayed here is the same as that previously displayed and shown in Figure 39. The congruence markings that have been added are displayed as accompanying text. Notice that, because we are combining written information with diagrammatic information here, we have made our system (slightly) heterogenous.

Finally, we can combine these two markings using the command `<c>ombine markings`. This command takes the place of transitivity: if we have a dseg or di-angle that is marked with two different markers, we can combine them into one marker that marks everything that is marked by either marking.

```
CDEG(1/1)% c
Type of marker to combine: (choices are <s>eg or <a>ng) s
Enter dseg:

Enter next seg index, or 0 to quit:15
Enter next seg index, or 0 to quit:0

CDEG(1/1)% v
marker2480 marks DSeg(solid1462) DSeg(solid1763)
```

```
DSeg(solid15)
```

Thus, we have shown how to construct an equilateral triangle on the given base, duplicating Euclid's Proposition 1.

Now, let's look at how **CDEG** handles a construction that results in an array of possibilities. We will look at the previously discussed construction shown in Figures 8 and 9. To get the starting diagram, we will load our saved diagram containing a single dseg, and add two new dots to it.

```
CDEG(1/1)% l
Enter file name: seg.cd
CDEG(1/1)% p
Diagram #1:
dot13 is surrounded by: region4 solid15
dot12 is surrounded by: region4 solid15
solid15 ends at dots dot12 and dot13
frame3 ends at loop in regions region4 and outerregion
dline14 is made up of dot13 solid15 dot12

region4 has boundry: frame3
        and contents:
Component #1: dot13 solid15 dot12 solid15

CDEG(1/1)% r
Enter region number: 4
CDEG(1/1)% r
Enter region number: 4
CDEG(1/1)% v
```

This diagram is shown in Figure 18 in Section 3.5.

After we connect the two new dots, the command prompt changes to indicate that the current diagram array contains 9 diagrams. We can look at each of these in turn by using the `se<t> pd` command, which controls which of the primitive diagrams in the array we are currently working with.

```
CDEG(1/1)% n
Enter first dot's number: 23
Enter second dot's number: 24
CDEG(1/9)% v
CDEG(1/9)% t
Enter pd number: 2
CDEG(2/9)% v
```

```
CDEG(2/9)% t
Enter pd number: 3
CDEG(3/9)% v
CDEG(3/9)% t
Enter pd number: 4
CDEG(4/9)% v
CDEG(4/9)% t
Enter pd number: 5
CDEG(5/9)% v
CDEG(5/9)% t 6
Enter pd number: 6
CDEG(6/9)% v
CDEG(6/9)% t
Enter pd number: 7
CDEG(7/9)% v
CDEG(7/9)% t 8
Enter pd number: 8
CDEG(8/9)% v
CDEG(8/9)% t
Enter pd number: 9
CDEG(9/9)% v
CDEG(9/9)% q
Are you sure you want to quit? yes
 Bye!
```

These diagrams are shown in Figures 19 and 20 in Section 3.5; they are the same diagrams that were shown in Figure 9.

References

Aaboe, Asger. 1964. *Episodes from Early Mathematics*. New York: Random House.

Anderson, Michael, Peter Cheng, and Volker Haarslev, eds. 2000. *Theory and Application of Diagrams*. Lecture Notes in Artificial Intelligence 1889. Berlin: Springer.

Barwise, Jon and Gerard Allwein, eds. 1996. *Logical Reasoning with Diagrams*. New York: Oxford University Press.

Bell, E. T. 1937. *Men of Mathematics*. New York: Simon & Schuster.

Blackwell, Allen, Kim Marriott, and Atsushi Shimojima, eds. 2004. *Diagrammatic Representation and Inference*. Lecture Notes in Artificial Intelligence 2980. Berlin: Springer.

Bourbaki, Nicolas. 1994. *Elements of the History of Mathematics*. Berlin: Springer-Verlag.

Dipert, Randall R. 2001. History of logic. In *Encyclopaedia Britannica*. `http://www.britannica.com`: Encyclopaedia Britannica, online edn.

Dodgson, Charles. 1885. *Euclid and His Modern Rivals*. London: MacMillan and Co., 2nd edn.

Emerson, Ralph Waldo. 1876. *Society and Solitude*. Boston: James R. Osgood and Company.

Euclid. 1956. *The Elements*. New York: Dover, 2nd edn. Translated with introduction and commentary by Thomas L. Heath.

Forder, Henry George. 1927. *The Foundations of Euclidean Geometry*. Cambridge: Cambridge University Press.

Gardner, Martin. 1982. *Logic Machines and Diagrams*. Chicago: University of Chicago Press.

Greaves, Mark. 2002. *The Philosophical Status of Diagrams*. Stanford: CSLI Publications.

Hammer, Eric. 1995. *Logic and Visual Information*. Stanford: CSLI Publications.

Hatcher, Allen. 2002. *Algebraic Topology*. Cambridge: Cambridge University Press. Freely available online at `http://www.math.cornell.edu/~hatcher/AT/ATpage.html`.

Hegarty, Mary, Bernd Meyer, and N. Hari Narayanan, eds. 2002. *Diagrammatic Representation and Inference*. Lecture Notes in Artificial Intelligence 2317. Berlin: Springer.

Hilbert, David. 1971. *Foundations of Geometry*. La Salle, Ill.: Open Court Publishing Co., 4th edn. Translated by Leo Unger.

Joseph, George Gheverghese. 1991. *The Crest of the Peacock: Non-European Roots of Mathematics*. London: I. B. Tauris & Co.

Kline, Morris. 1967. *Mathematics for the Nonmathematician*. New York: Dover Publications.

Kline, Morris. 1980. *Mathematics: The Loss of Certainty*. New York: Oxford University Press.

Luengo, Isabel. 1995. *Diagrams In Geometry*. Ph.D. thesis, Indiana University.

Luengo, Isabel. 1996. A Diagrammatic Subsystem of Hilbert's Geometry. In G. Allwein and J. Barwise, eds., *Logical Reasoning with Diagrams*. New York: Oxford University Press.

Manders, Kenneth. 1995. The Euclidean diagram. Unpublished manuscript, draft dated 5/95.

Miller, Nathaniel. 2006. Computational complexity of diagram satisfaction in Euclidean geometry. *Journal of Complexity* 22(2):250–274.

Neugebauer, Otto and A. Sachs. 1945. *Mathematical Cuneiform Texts*. Lancaster, PA: Lancaster Press.

Peirce, Charles Sanders. 1960. *Collected Papers*, vol. IV. Cambridge: Belknap Press. Edited by Charles Hartshorne and Paul Weiss.

Plutarch. 1878. The Symposiacs, Book XIII, Question II. In *Plutarch's Morals*, vol. III. Boston: Little, Brown, and Co. Translated from the Greek by Several Hands. Corrected and Revised by William W. Goodwin, with an Introduction by Ralph Waldo Emerson.

Plutarch. 75. Parallel lives: Marcellus. Available online at `http://classics.mit.edu/Plutarch/marcellu.html`.

Renegar, James. 1992. On the computational complexity and geometry of the first order theory of the reals. *Journal of Symbolic Computation* 13(3):255–352.

Shin, Sun-Joo. 1994. *The Logical Status of Diagrams*. Cambridge: Cambridge University Press.

Simmons, George F. 1985. *Calculus with Analytic Geometry*. New York: McGraw Hill.

Tarski, Alfred. 1951. *A Decision Method for Elementary Algebra and Geometry*. Berkeley, CA: University of California Press.

Whitehead, Alfred North. 1911. *An Introduction to Mathematics*. London: Williams and Norgate.

Index