Solving Knights-and-Knaves with one equation*

Francesco Ciraulo Samuele Maschio
Department of Mathematics, University of Padova
{ciraulo,maschio}@math.unipd.it

Logic puzzles are often solved without any specific mathematical tool. By fully exploiting the deep connection between Logic (Boolean algebras) and Commutative Algebra (Boolean rings), we show here that Raymond Smullyan's well-known Knights-and-Knaves puzzles, as well as other variations, can often be solved by means of a single algebraic equation.

1 Introduction

Among the many uses of the word *logic* in everyday life (admissions tests, escape rooms and so on), Knights-and-Knaves puzzles are rare examples of something which is commonly called *logic* and that, no doubt, *logic* is indeed. So much has been said about this brilliant creation[1] of Raymond Smullyan's (see [6]) that it is difficult to believe there can be something really new and interesting on that topic nowadays. Our aim is to prove that this is wrong!

As everybody knows, only two breeds live in Smullyan's island: Knights, who *always* say the truth, and Knaves, who *always* lie. A typical puzzle goes like this: some inhabitants of the island state something, usually about their own or others' nature (which they know very well). The challenge is to guess which is their nature. Here comes an example: *First says that Third is a knave; Second confirms that and, in addition, he also says that First and Third do not share the same nature (of course). Which is the nature of the three inhabitants?* The authors want to convince the readers that solving that puzzle has the same complexity as solving the following system of equations:

$$\begin{cases} a = c + 1 \\ b = c + 1 \\ b = a + c \end{cases}$$

(which makes solving the puzzle a well-determined mathematical task, rather

*This is the preprint version of the paper published in The College Mathematics Journal 51 (2020) pp. 82-89 (www.tandfonline.com/doi/full/10.1080/07468342.2020.1698931).

[1] To be honest, the first appearance of a puzzle of this kind is in [4], where the two types of inhabitants are called the Arbus and the Bosnins.

than a case analysis that can easily end in an ungovernable bunch of possible roads, at least in more complex examples).

2 Logic via Algebra

2.1 Propositions and their Truth Values

According to the Oxford Online Dictionary, a *proposition*, as meant by logicians, is a "statement that expresses a concept that can be true or false" although, we add, we need not know which of the two alternatives holds.[2] Here are some examples:

- "Two and two is four."
- "One and one is zero."
- "ETs exist, and they pilot UFOs. "

For convenience, we will use 0 and 1 for the *truth values* "false" and "true" respectively. If P is a proposition, its truth value will be written $[\![P]\!]$. Then

- $[\![P]\!] = 1$ if and only if P is true;
- $[\![P]\!] = 0$ if and only if P is false.

Note that expressions like "$[\![P]\!] = 1$" or "$[\![P]\!] = 0$" are propositions as well.[3] So it makes sense to wonder what their truth value is... Clearly "$[\![P]\!] = 1$" has the same truth value as P, that is

$$[\![\text{“}[\![P]\!] = 1\text{”}]\!] = [\![P]\!]$$

and then "$[\![P]\!] = 1$" is equivalent to P. The general rule is that

$$[\![p = 1]\!] = p \tag{1}$$

for every p in $\{0, 1\}$. (To be pedantic we should have written $[\![\text{“}p = 1\text{”}]\!] = p$, with quotation marks, but we are not pedantic!)

Up to equivalence, there are just two possible propositions: "1=1" and "0=1". However, we will often deal with propositions which depend on some variable (these are traditionally called "propositional functions", but we prefer to call them simply "propositions"). So, more generally, the truth value of a proposition $P(x)$, with x ranging over some set X, is a Boolean value $p(x)$ depending on x, that is, a function $p : X \to \{0, 1\}$; and $P(x)$ is equivalent to "$p(x) = 1$".[4]

[2] This conclusion is characteristic of the so-called classical logic: the twentieth century saw the emergence of a number of alternative logics in which the truth values are more than two, or even infinite, or are not explicitly determined, or can coexist, or ....

[3] This makes sense because we are working in our informal, natural language (which makes the notion of proposition quite open). If, on the contrary, propositions are identified with formulas over a given formal language, then statements like "$[\![P]\!] = 1$" belong to the meta-language (whatever that means) and need not correspond to formulae of the object language.

[4] So, up to equivalence, a proposition is something of the form "$p(x) = 1$", where p :

2.2 Modular arithmetic and logical connectives

Writing 0 for 'false' and 1 for 'true' is a trick that can be traced back to the work of George Boole [1, 2].[5]

CONNECTIVE	LATTICE form	RING form
conjunction (AND)	$p \wedge q$	pq
exclusive disjunction (XOR)	$(p \vee q) \wedge \neg(p \wedge q)$	$p + q$
inclusive disjunction (OR)	$p \vee q$	$pq + p + q$
negation	$\neg p$	$p + 1$
implication ($\rightarrow$)	$q \vee \neg p$	$pq + p + 1$
biconditional ($\leftrightarrow$)	$(p \wedge q) \vee (\neg p \wedge \neg q)$	$p + q + 1$
Sheffer stroke (NAND)	$\neg(p \wedge q)$	$pq + 1$
Pierce's arrow (NOR)	$\neg(p \vee q)$	$pq + p + q + 1$

Table 1: Boolean connectives in the language of Boolean rings

Note that $p + q + 1 = 1$ if and only if $p = q$. Thus, $p + q + 1$, besides being the truth value of the bi-implication between P and Q is also the truth value of the proposition "$p = q$". In other words, we have the following generalization of equation (1)

$$[\![p = q]\!] = p + q + 1 \ . \tag{2}$$

Quantifiers could be treated in a similar way too. Let P be a proposition depending on a variable x ranging over some set D (and possibly depending on other variables). This means that the truth value p of P is a boolean valued function of the variable $x \in D$ (and possibly of other variables). Then the truth value of the universal proposition $(\forall x \in D)P(x)$ is simply the minimum $\min_{x \in D} p(x)$ of all $p(x)$ for $x \in D$. And the truth values of $(\exists x \in D)P(x)$ is $\max_{x \in D} p(x) = 1 + \min_{x \in D}(1 + p(x))$.

3 One puzzle, one equation

The basic fact one has to keep in mind in order to understand Smullyan's creatures is the following.

Fact 1. If an inhabitant A says "P", then the information we get is that

$$a = p$$

where $a = [\![$"A is a knight"$]\!]$ and $p = [\![P]\!]$.[6]

$X \rightarrow \{0, 1\}$ for some set X. This could serve as a mathematical definition of the notion of a proposition (this is essentially the standard approach in categorical logic).

[5] Even though he did not work with modular but rather with standard arithmetic

[6] The same information can be coded by saying that $1 = (a \wedge p) \vee (\neg a \wedge \neg p)$. You must agree that $a = p$ is much clearer.

A typical puzzle contains a stock of statements $P_1 ... P_n$ made by some inhabitants $A_1, .., A_n$, respectively, and hence can be mathematized as a system of equations of the following form

$$\begin{cases} a_1 = p_1 \\ \dots \\ a_n = p_n \end{cases}$$

Every such a system can be reduced to a single equation, namely

$$\Pi_{i=1}^{n}(p_i + a_i + 1) = 1$$

although this is seldom convenient.[7]

3.1 Examples

In the following examples we will always use upper case letters A, B,... to denote the inhabitants of Smullyan's island. Lower case letters a, b, ... will denote the truth values of the propositions "A is a knight", "B is a knight",..., respectively.

Example 1. A says "Both B and I are knaves".
Solution. The corresponding equation is $a = (b+1)(a+1)$, that is, $0 = ab+b+1$, which can be written as $(a+1)b = 1$. So $a + 1 = 1 = b$ (A is a knave and B is a knight).[8]

Example 2. A says "If B were a knave, then I'd be a knave too".
Solution. The corresponding equation is $a = (b + 1)(a + 1) + (b + 1) + 1$, that is, $ab = 1$; so $a = b = 1$.

Example 3. (This is essentially the "father" puzzle of [4].) We met three inhabitants and we asked A which breed is s/he. Unfortunately, we did not catch what s/he said. Luckily, B and C are going to help us. So B tells us that A said to be a knight. On the contrary, C tells us that A said to be a knave.
Solution. The corresponding equations are $b = [\![A \text{ says that } A \text{ is a knight}]\!] = [\![a = a]\!] = 1$ and $c = [\![a = a + 1]\!] = 0$. (Obviously, the only thing A could have said is to be a knight.)

Example 4. A and B said something about their reciprocal nature without our hearing it. Then C tells us "A said that B is a knave and B said that A is a knight."
Solution. $c = [\![a = b+1]\!] \cdot [\![b = a]\!] = (a+(b+1)+1)(b+a+1) = (a+b)(a+b+1) = 0$. So our information amounts precisely to $c = 0$.

[7] Note, for completeness, that inequalities in $\{0, 1\}$ also reduce to equations: $p \leq q$ means the same as $pq = p$. By the way, as a little exercise in logic, the reader is invited to think about the difference between the proposition $P \to Q$ and the inequality $[\![P]\!] \leq [\![Q]\!]$. And what about the proposition "$[\![P]\!] \leq [\![Q]\!]$"?!

[8] Note the advantage of the ring notation. In lattice notation, instead, we should have solved the equation $a = \neg b \wedge \neg a$. How would one react in front of it? Case analysis!

Example 5. This is what four inhabitants told us:

A : "There is at least one knave among us."

B : "There are at most two knights among us."

C : "There are at least three knaves among us."

D : "There are no knights among us."

Solution. Conditions involving cardinality are usually quite complex: hence we look for a shortcut. The first equation is $a = 1 + abcd$ ("We are not all knaves."), that is, $a(1 + bcd) = 1$; so $a = 1$ and $bcd = 0$. The last equation is $d = (a + 1)(b + 1)(c + 1)(d + 1)$ from which, given $a = 1$, we obtain $d = 0$. Therefore the first and the last equation together says precisely that $a = 1$ and $d = 0$. The third equation, given that $a = 1$ and $d = 0$, must be equivalent to $c = (b+1)(c+1)$, that is, $b(c+1) = 1$; so $b = 1$ and $c = 0$.[9] These values satisfy the second equation as well.

4 Each question, the right... question!

Each proposition P is naturally associated to the question "Which is the truth value of P?" which, for the sake of readability, we simply write "P?". We can read it as the Yes/No question "Is P true?" provided that we identify "Yes" with 1, and "No" with 0.

Let us write $r(P, A)$ for the answer that an inhabitant A would give to the question "P?". If we can assume that A knows about $[\![P]\!]$ (and that s/he is willing to answer our question), then we have $r(P, A) = p + a + 1$ because $r(P, A)$ must be p if and only if A is a knight. As we know, $r(P, A)$ is the truth value of the proposition "$r(P, A) = 1$", which can be read as follows "*If asked about, A would answer that P is true.*" (A could assert P). Let us write $R(P, A)$ for such a proposition, so that $[\![R(P, A)]\!] = r(P, A)$. Summing up,

Fact 2. Let $R(P, A)$ be the proposition "If asked about, A would answer that P is true." (equivalently, "A could affirm P."). Then

$$r(P, A) = [\![R(P, A)]\!] = p + a + 1 \tag{3}$$

where, as usual, $p = [\![P]\!]$ and $a = [\![$"A is a knight"$]\!]$.

Now if we ask B "$R(P, A)$?", then B answers will be $r(R(P, A), B)$, that is, $p + a + b$. In particular, if B is A, then

$$r(R(P, A), A) = p \tag{4}$$

[9] Note that a condition of the form $x = (x+1)y$ gives us complete information about x and y, namely $x = 0 = y$.

So A's answer to the question "$R(P, A)$?" is the correct truth value of P, regardless of the nature of A.[10] We can therefore formulate the following principle (compare with the *embedded question lemma* in [5]).

Fact 3. In order to discover the correct truth value of a certain proposition P, it is enough to ask any inhabitant the following question:

If I asked you "P?", would you answer "Yes"? .

The answer we get is "Yes" if and only if $[\![P]\!]$=1.

Example 6. You're facing two doors; one hides a treasure, the other a goat.The doors are guarded by two guards; one is a knight, the other is a knave (and you don't know which, of course). You are allowed to ask just one of them a yes/no question. What shall you ask?
Solution. I'll ask any of the two guards the following question "*If I asked you "Does your door hide the treasure?", would you answer "Yes"?*", and I apply the previous fact.

5 Variations on a theme

The island of knights and knaves is only the simplest of a variety of more and more complex situations. For instance, what happens if a further breed of people exists, call them *pages*, which tell the truth every other day? In this extended context, a typical problem is to establish the nature of some inhabitants by having information on what they have said in two consecutive days. To formalize the problem, we can associate to each inhabitant A a pair (a_1, a_2) of truth values: a_i is 1 if and only if A is telling the truth on the i-th day. So

- A is a knight iff $(a_1, a_2) = (1, 1)$ iff $a_1 a_2 = 1$;
- A is a knave iff $(a_1, a_2) = (0, 0)$ iff $(1 + a_1)(1 + a_2) = 1$;
- A is a page iff either $(a_1, a_2) = (1, 0)$ or $(a_1, a_2) = (0, 1)$ iff $a_1 + a_2 = 1$.

Example 7. Yesterday I met A, B and C: A said 'I'm a page'; B said 'I'm a knight'; C said 'I'm a knave'. This morning I met them again: A said 'B is a knight'; B said 'C is a page'; C said 'A is a page'. Who is who?
Solution. The problem data are represented in the following system

$$\begin{cases} a_1 = a_1 + a_2 \\ b_1 = b_1 b_2 \\ c_1 = (1 + c_1)(1 + c_2) \\ a_2 = b_1 b_2 \\ b_2 = c_1 + c_2 \\ c_2 = a_1 + a_2 \end{cases}$$

[10] Note that asking A the question "$R(P, A)$?" is like asking him/her the following question: '*If I asked you "P?", would you answer "Yes"?*'.

which gives $a_2 = b_1 = c_1 = 0$, $c_2 = b_2 = a_1 = 1$. So they are all pages.

5.1 Knights, knaves, and normals

Normal people sometimes lie and sometimes tell the truth. To each inhabitant A let us associate two Boolean values, namely $a = [\![A \text{ is a knight}]\!]$ and $a' = [\![A \text{ is a knave}]\!]$, on the proviso that $aa' = 0$ (A cannot be both a knight and a knave at the same time). So

- A is a knight iff $a = 1$;
- A is a knave iff $a' = 1$;
- A is normal iff $a + a' = 0$.

Assume that A says P; what we know is precisely that the two implications (A is a knight)$\to P$, and (A is a knave)$\to \neg P$ must be true. These data can be represented in one of the following equivalent forms (where $p = [\![P]\!]$):

$$\begin{cases} ap + a + 1 = 1 \\ a'(p+1) + a' + 1 = 1 \end{cases} \qquad \begin{cases} a = ap \\ a'p = 0 \end{cases} \qquad a = (a + a')p$$

(note that the condition $aa' = 0$ follows: if $a = 1 = a'$, then $0 = p = 1$).

Example 8 ([6], puzzle 40)**.** Given that A says 'B is a knight', and B says 'A is not a knight', prove that at least one of them is telling the truth, although s/he is not a knight.
Solution. The problem corresponds to the system displayed below on the left, which is equivalent to that on the right.

$$\begin{cases} a = ab \\ a'b = 0 \\ b = b(1+a) \\ b'(1+a) = 0 \end{cases} \qquad \begin{cases} a = 0 \\ a'b = 0 \\ b' = 0 \end{cases}$$

We must prove that $[(a = 0) \wedge (b = 1)] \vee [(b = 0) \wedge (a = 0)]$, which is obvious because we know that $a = 0$ is true, and $(b = 1) \vee (b = 0)$ is always true. (Compare this solution to the one proposed in [6].) Of course, we know more than that: of the nine possible scenarios, we are left with three: (i) A is a knave and B is normal, (ii) A is normal and B is a knight, (iii) A and B are both normals.

The hardest logic puzzle ever [3]. We meet A, B and C. We already know that one of them is a knight, one is a knave and one is normal.[11] Unfortunately,

[11] Assume that normals have two mental states which determine whether they are going to speaks truly or falsely. They persist in the same state until they say something; then their state can (but need not) switch. Knights and knaves can be seen as particular cases: they always persist in the same state.

they cannot speak our language, tough they understand it. We know that they use "ja" and "da" to mean "yes" and "no" but, of course, we do not know which means which. Our task is to discover their identity by asking them no more than three questions; and each question has to be put to one of them.
Solution. The simplest solution is given in [5]; here we simply formalize it. We use the following variables

$$\begin{aligned} a &= [\![\text{In her/his current state, } A \text{ would speak truly}]\!] \\ b &= [\![\text{In her/his current state, } B \text{ would speak truly}]\!] \\ c &= [\![\text{In her/his current state, } C \text{ would speak truly}]\!] \\ j &= [\![\text{'ja' means 'yes'}]\!] \end{aligned}$$

(it is understood that a, b and c vary with time, even though this has not been explicitly indicated). As usual, given any proposition P, let p be $[\![P]\!]$.
We are going to adapt the method presented in section 4. For $X \in \{A, B, C\}$, let us define

$$r_j(P, X) = \begin{cases} 1 \text{ if (in her/his current state) } X\text{'s answer to } P\text{? is '}ya\text{'}; \\ 0 \text{ if (in her/his current state) } X\text{'s answer to } P\text{? is '}da\text{'}. \end{cases}$$

So $r_j(P, X) = r(P, X)$, as defined in section 4, if and only if 'ja' means 'yes'. In other words, $r_j(P, X) = p + x + j$. Let $R_j(P, X)$ be the proposition "$r_j(P, X) = 1$" as usual. Then the proposition $R_j(R_j(P, X), X)$ depends neither on X nor on j, as its truth value is $(p + x + j) + x + j$, that is, p.
Summing up we have the following "embedded question lemma" [5].

Fact 4. Let $E(P)$ be $R_j(R_j(P, X), X)$. Then X's answer (in her/his current state) to the question $E(P)$? is the correct truth value of P, provided that 'ya' is interpreted as 1 and 'da' as 0.

So $E(P)$ is equivalent to the following proposition:

X answers 'ja' to the question "Would your answer 'ja' to the question P? ?"

So asking X the question E(P)? is like asking her/him:

In case I asked you
"Would you answer 'ja' to the question P? ?",
would you answer 'ja'?

This is enough to solve the puzzle, of course, since we can get three correct pieces of information about the nature of A, B and C. For instance, we can use our first question to discover whether A is a knight or not. In the former case, we use the second question to understand whether B is a knave or a normal; and we are done. In the latter case, we use the second question to decide whether A is a knave or a normal; then we need the last question to reveal B's (and hence C's) nature.

References

[1] George Boole. The Mathematical Analysis of Logic, Being an Essay Towards a Calculus of Deductive Reasoning. Philosophical Library, New York, N. Y., 1948.

[2] George Boole. An investigation of the laws of thought, on which are founded the mathematical theories of logic and probabilities. Dover Publications, Inc., New York, 1957.

[3] Boolos, G. 1996. The hardest logic puzzle ever. The Harvard Review of Philosophy 6: 62–65. Repr. in his Logic, Logic, and Logic, 406–10. 1998. Cambridge, Mass.: Harvard University Press.

[4] Maurice Kraitchik. Mathematical Recreations. Dover, 1953.

[5] Brian Rabern and Landon Rabern. A simple solution to the hardest logic puzzle ever. Analysis, 68(2):105-112, 2008.

[6] Raymond Smullyan. What is the Name of this Book? Prentice-Hall, 1978.

[7] M. H. Stone. The theory of representations for Boolean algebras. Trans. Amer. Math. Soc., 40(1):37-111, 1936.