The Analysis Behind The Intuition

Sarah Dennis

October 29, 2019

1 Introduction

Often in Mathematics we encounter concepts that are highly intuitive at a broad level, but that need an intense amount of detail and rigour to make their construction fit for mathematical use. Take the derivative for an example. Intuitively, the derivative is the slope, and slope is not a difficult concept to grasp. We can relate the slope of a function to the slope of a hill or a road – sometimes steeper and sometimes flatter. But the construction of the derivative itself is quite complex. In high school algebra many are taught that 'functions' have the form $y = mx + b$ where m is the slope. No one would dare tell you that $y' = m$ and this is *why* we call m the slope. The derivative is complicated because outside of these linear functions, it is not sufficient to say 'the derivative of a function is the slope'. We need to take the slope at every point – but this has *no* inherent meaning because a point has no slope! So to take the derivative we need to involve a limit, which then requires we consider continuity. Again we encounter an intuitive understanding for continuity. That is, a function f is continuous if values that are *close* in the domain, get mapped by f to values that are *close enough* in the codomain. But what do we mean by close and close enough? We bring ϵ and δ proofs into the mix and suddenly the idea of continuity is very rigorous, but also very complicated.

In this paper we will work through proofs and examples related to Lebesgue measure and integration, Baire category theory, Euler's sum, and the gamma function. These topics follow the same trend whereby the broad concepts are not difficult to grasp, but we need a lot of ground work to give enough detail for precise and accurate definitions.

Intuitively, sets that have measure zero are very small. But like values that are *close enough*, how small is *very small*? Lebesgue shows us that we can still use Riemann integration so long as there are not *too many* discontinuities. But how many is too many? Turns out that any more than a *very small* amount is too many. Likewise, trigonometric functions such as sine and cosine are intuitive. A student in a basic calculus class could tell you that sine is a smooth wave function that crosses 0 at 0, and gosh there is something to do with $\pi \ldots$ Hopefully this calculus student would also confirm that we can write a polynomial as a product of its roots. So in the same way we will approximate $\sin(x)$ as a product of its roots – that pesky π again! Finally, we know e^x as the nice function whose derivative is itself and we know $x!$ is the nice function where $(x+1)! = (x+1)x!$. But what is a *nice* function, how did someone come up with e, and what if we want the factorial to be defined for all numbers? Our intuitive understanding of these concepts is not enough for us to perform the calculations we want. So we will demonstrate here some of the messy mechanics that is hidden underneath some of mathematics' more beautiful and seemingly simple ideas.

2 Lebesgue's Criterion For Riemann Integrability

2.1 Riemann Integrability

The Riemann integral is that which we typically encounter in introductory calculus. We have a function f defined on an interval $[a, b]$. We construct some partition $P = \{x_0, x_1, ...x_n\}$ of this interval, and form rectangles with base $[x_i, x_{i+1}]$ and height $f(c)$ for some $c \in [x_i, x_{i+1}]$. The area of each rectangle is of course *base* $\times$ *height*; that is, we have area $f(c)(x_i - x_{i+1})$. The total area under the curve is the (Riemann) sum of all these areas.

$$\sum_{i=0}^{n-1} f(c)(x_i - x_{i+1}) \tag{1}$$

The finer our partition, the more accurate this measurement of area will be. Limits and the axiom of completeness all enable us to construct partitions that are 'fine enough' for an accurate measurement.

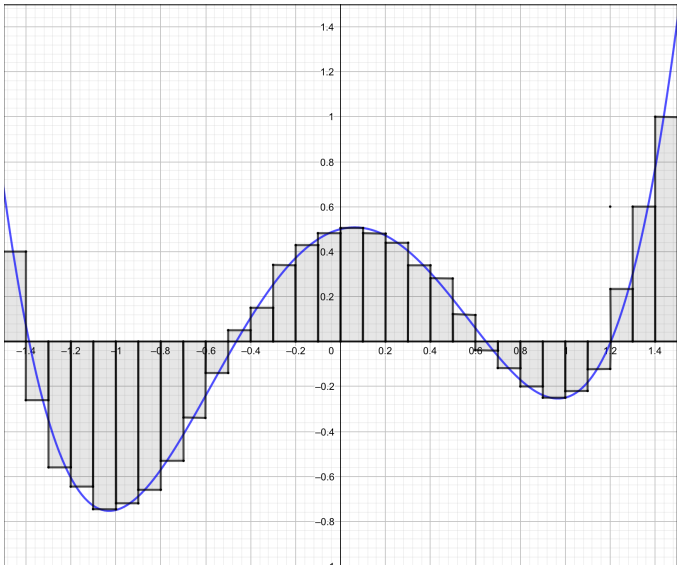


Figure 1: Reimann integral for $f(x) = x^4 - 2x^2 + \frac{1}{4}x + \frac{1}{2}$

We then have the following theorem.

Theorem 2.1. *If f is continuous on $[a, b]$, then f is Riemann-integrable on $[a, b]$.*

So what can do with functions that are not continuous?

2.2 Lebesgue Measure

In general, we place a measure on a set to conceptualise how large that set is. This becomes challenging when dealing with subsets of the uncountable real numbers, especially if these subsets are defined in terms of intervals. We restrict our interest for now to sets that have measure zero.

Definition 2.1. *A set $A \subseteq \mathbb{R}$ has measure zero if, for all $\epsilon > 0$, there exists a countable collection of intervals O_n such that*

$$A \subseteq \bigcup_{n=1}^{\infty} O_n \qquad \text{and} \qquad \sum_{n=1}^{\infty} |O_n| < \epsilon.$$

Exercise 7.6.3. Show that any countable set has measure zero.

Proof. Let X be a countable set with bijection $f : X \to \mathbb{N}$ and take $\epsilon > 0$ to be an arbitrary real number. Let $x_n = f(x)$ for $x \in X$ and set

$$O_n = (x_n - \frac{\epsilon}{2^{n+2}}\ ,\ x_n + \frac{\epsilon}{2^{n+2}}). \tag{2}$$

Then $\{O_n\}_{n\in\mathbb{N}}$ is a countable collection of open intervals that covers X. Furthermore, for any $n \in \mathbb{N}$ the size of the open interval O_n is given by,

$$|O_n| = \frac{2\epsilon}{2^{n+2}} = \frac{\epsilon}{2^{n+1}}. \tag{3}$$

So the sum of the lengths of all O_n is

$$\sum_{n=1}^{\infty} |O_n| = \sum_{n=1}^{\infty} \frac{\epsilon}{2^{n+1}} = \frac{\epsilon}{2} \sum_{n=1}^{\infty} \frac{1}{2^n} = \frac{\epsilon}{2} < \epsilon \tag{4}$$

Thus $\{O_n\}_n$ satisfies the sufficient properties to declare that X has measure zero. □

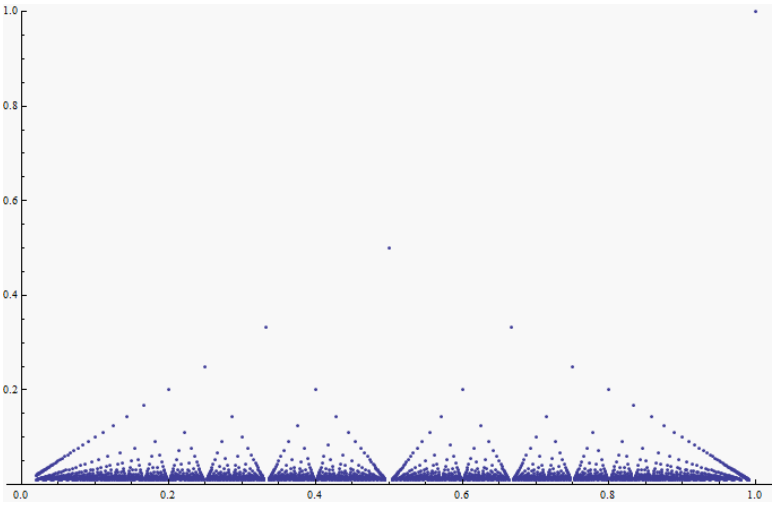


Figure 2: The Thomae Function

$$T(x) = \begin{cases} \frac{1}{q} & \text{if } x = \frac{p}{q} \text{ for } p, q \in \mathbb{Z} \text{ relatively prime, } q \neq 0 \\ 0 & \text{if } x \text{ is irrational} \end{cases} \tag{5}$$

The range of the Thomae function (5) has measure zero as it is a set of only rational points, and the rational numbers are countable.

2.2.1 The Cantor Set

The Cantor set is a subset of the unit interval formed by repeatedly removing the middle third of the intervals that remain. What is left over? We cannot have any intervals left in the Cantor set because we would have to have its middle third removed. But certainly the end points of every interval remain, that is, $C = \{0, 1, \frac{1}{3}, \frac{2}{3}, \frac{1}{9}, ...\}$. The Cantor set is often described as dust. And as we will now see, this is not an inaccurate description.

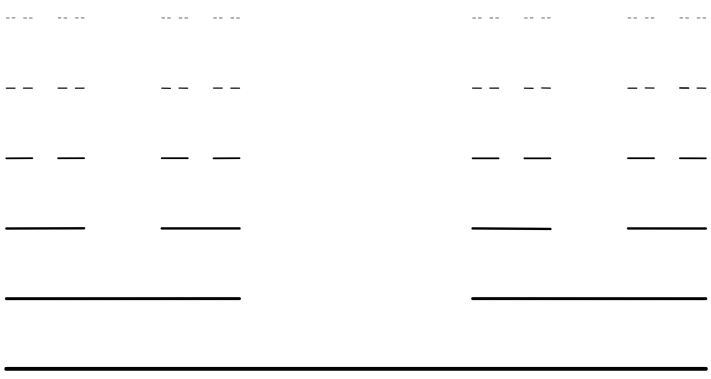


Figure 3: The Cantor Set

Exercise 7.6.4. Prove that the Cantor set has measure zero.

Proof. Denote the Cantor set by C, and at the n^{th} step in the process of removing the middle thirds of the unit interval, denote the set of remaining elements by C_n. Observe that after any step n, the total length removed from the unit interval is $\sum_{n=1}^{\infty} \frac{2^{n-1}}{3^n}$. This geometric series is converging to 1. So let $\epsilon > 0$ and take $N \in \mathbb{N}$ large enough for

$$\sum_{n=1}^{N} \frac{2^{n-1}}{3^n} > 1 - \epsilon. \tag{6}$$

That is, proceed to step N where we have that the length removed is greater than $1 - \epsilon$ for any $\epsilon > 0$. Now define the collection of open intervals removed at step N as $\{O_k\}_{k \in \{1,\ldots,2^{N-1}\}}$. The complement of this collection, $[0,1] - \cup_k \{O_k\}$, is then a closed cover for C_N and the total underlying set C. Furthermore, the total length of this cover is

$$|(\cup_k \{O_k\})| = \sum_{n=1}^{N} \frac{2^{n-1}}{3^n} > 1 - \epsilon. \tag{7}$$

Note that the length of the unit interval is 1. Which gives that the length of the complement is

$$|[0,1] - (\cup_k \{O_k\})| = 1 - |(\cup_k \{O_k\})| < 1 - (1 - \epsilon). \tag{8}$$

It follows then that

$$|[0,1] - (\cup_k \{O_k\})| < \epsilon. \tag{9}$$

Take the smallest open set containing the complement of $\cup_k \{O_k\}$. Then we will have an open cover of total length less than or equal to ϵ. Thus the Cantor set has measure zero. □

Exercise 7.6.5. Show that if two sets A and B each have measure zero, then $A \cup B$ has measure zero as well. In addition, discuss the proof of the stronger statement that the countable union of sets of measure zero also has measure zero.

Proof. Let A and B be sets of measure zero. Then we have

$$A \subseteq \bigcup_{i=1}^{\infty} O_i \text{ and } B \subseteq \bigcup_{j=1}^{\infty} P_j \tag{10}$$

for collections of open sets $\{O_i\}_{i \in \mathbb{N}}$ and $\{P_j\}_{j \in \mathbb{N}}$ where $\sum_i |O_i| \leq \epsilon$ and $\sum_j |P_j| \leq \epsilon$ for any $\epsilon > 0$. Set $\epsilon_1 = \frac{\epsilon}{2}$ and take

$$\{Q_k\}_{k \in \mathbb{N}} = \{O_i\}_i \cup \{P_j\}_j. \tag{11}$$

The countable union of countable sets is countable, so $\{Q_k\}_k$ is certainly a countable collection. Furthermore, we have $A \cup B \subseteq \cup_k Q_k$. Finally, since $\sum_i |O_i| \leq \epsilon_1$ and $\sum_j |P_j| \leq \epsilon_1$ we have

$$\sum_{k=1}^{\infty} |O_k| \leq \sum_{i=1}^{\infty} |O_i| + \sum_{j=1}^{\infty} |P_j| \leq \epsilon_1 + \epsilon_1 = \epsilon. \tag{12}$$

Thus, we have that $A \cup B$ also has measure zero. □

To show that a countable union of open sets will have measure zero will be a similar process. We can define a countable collection of open sets as the union of the countable covers of all the sets in the union. The double summation will come in to play in the equivalent of equation 10 where we are summing the lengths. To show that the new cover has length less than epsilon, we need each cover in the union to have length less than the a fraction of epsilon (given by the number of covers in the union). All of this being done will show that a countable union of sets with measure zero also has measure zero.

2.3 Measure Zero and Integrability

2.3.1 α-continuity

The following definitions of α-continuity are the foundation for the punchline of Lebesque's criterion for Riemann integrability.

Definition 2.2. *Let f be defined on $[a,b]$, and let $\alpha > 0$. The function f is α-continuous at $x \in [a,b]$ if there exists $\delta > 0$ such that for all $y, z \in (x-\delta, x+\delta)$ it follows that $|f(y) - f(z)| < \alpha$.*

Definition 2.3. *Let f be a bounded function on $[a,b]$. For each $\alpha > 0$, define D^α to be the set of points in $[a,b]$ where the function f fails to be α-continuous.*

$$D^\alpha = \{x \in [a,b] : f \text{ is not } \alpha\text{-continuous at } x\}. \tag{13}$$

Now, let

$$D = \{x \in [a,b] : f \text{ is not continuous at } x\}. \tag{14}$$

2.3.2 Criterion for Integrability

Recall that previously we knew that f continuous on $[a,b]$ implied that f Riemann integrable on $[a,b]$. We are now ready to go a step further and slightly loosen the restriction that of f continuous.

Theorem 2.2. *Let f be a bounded function defined on the interval $[a,b]$. The f is Riemann-integrable if and only if the set of points where f is not continuous has measure zero.*

This leads to some surprising results. We now know that in fact the Thomae function (5), which is discontinuous at all rational points, is in fact Riemann Integrable because this set of discontinuities has measure zero. The integral of the Thomae function on any closed interval will be equal to 0. We also have the interesting example of the Volterra function which is everywhere differentiable but whose derivative (having discontinuities at uncountably many points) has a non-Riemann-integrable derivative.

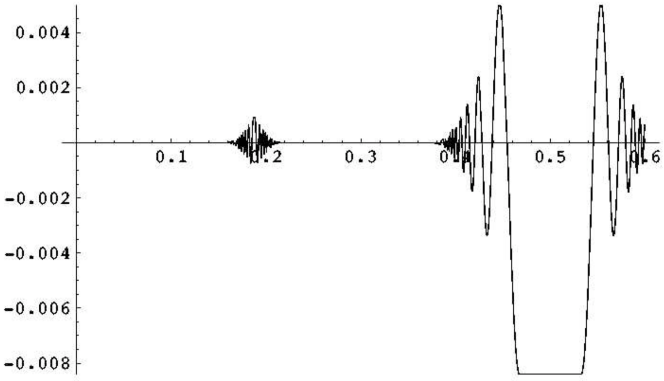


Figure 4: The Volterra Function [2]

Exercise 7.6.13.

a) Show that if f and g are integrable on $[a,b]$, then so is the product fg.

b) Show that if g is integrable on $[a.b]$ and f is continuous on the range of g, then the composition $f \circ g$ is integrable on $[a,b]$.

Proof. We will begin by showing function composition, so we can use this fact for showing function product.

b) Let g be a function integrable on $[a,b]$, and let f be continuous on the range of g. First suppose that f is continuous on $[a,b]$. Then certainly $f \circ g$ is continuous on $[a,b]$. So $D_{f\circ g}$ has measure zero and $f \circ g$ is

Riemann integrable on $[a, b]$. Now suppose instead that f has finitely many discontinuities on $[a, b]$. Then D_f is countable, so D_f has measure zero. The set of points where $f \circ g$ is not continuous is a subset of where f is not continuous. So $D_{f\circ g}$ is also countable and has measure zero. Thus $f \circ g$ is integrable on $[a, b]$.

a) Now suppose f and g are integrable on $[a, b]$ and observe,

$$fg = \frac{1}{2}[(f+g)^2 - f^2 - g^2] \tag{15}$$

Define $h(x) = x^2$. Then since $f^2 = h \circ f$, and f and h are continuous, f^2 is Riemann integrable, (and similarly for g^2). Thus we have that each component above is Riemann integrable on $[a, b]$, so their sum is also integrable on $[a.b]$. □

3 The Lebesgue Integral

With the new understanding that we can integrate functions so long as their discontinuities form a set of measure zero, we refine our concept of integrability so as to no longer depend on continuity. The Lebesgue method of integration has become standard, and will be no different to Riemann-integration on continuous functions.

Definition 3.1. *Let*

$$P = \{x_0, x_1, x_2, ..., x_n\} \tag{16}$$

be a partition of $[a, b]$. *A tagged partition is one where in addition to* P *we have chosen points* c_k *in each of the subintervals* $[x_{k-1}, x_k]$ *Then, given a function* $f : [a, b] \to \mathbb{R}$ *and a tagged partition* $(P, \{c_k\}_{k=1}^n)$, *the Riemann sum generated by this partition is given by*

$$F(f, p) = \sum_{k=1}^{n} f(c_k)(x_k - x_{k-1}). \tag{17}$$

Also recall we define an upper sum as

$$U(f, p) = \sum_{k=1}^{n} M_k(x_k - x_{k-1})$$

where

$$M_k = \sup\{f(x) : x \in [x_{k-1}, x_k]\},$$

and the lower sum as

$$L(f, p) = \sum_{k=1}^{n} m_k(x_k - x_{k-1}) \tag{18}$$

where

$$m_k = \inf\{f(x) : x \in [x_{k-1}, x_k]\}. \tag{19}$$

Definition 3.2. *Let* $\delta > 0$. *A partition* P *is* δ*-fine if every subinterval* $[x_{k-1}, x_k]$ *satisfies* $x_k - x_{k-1} < \delta$.

Theorem 3.1. *A bounded function* $f : [a, b] \to \mathbb{R}$ *is Riemann-integrable with* $\int_a^b f = A$ *if and only if, for ever* $\epsilon > 0$, *there exists a* $\delta > 0$ *such that, for any tagged partition* $(P, \{c_k\})$ *that is* δ*-fine it follows that*

$$|R(f, P) - A| < \epsilon \tag{20}$$

Exercise 8.1.4. a) Show that if f is continuous, then it is possible to pick tags $\{c_k\}_{k=1}^n$ so. that

$$R(f, P) = U(f, P). \tag{21}$$

Similarly, there are tags for which $R(f, P) = L(f, P)$ as well.

b) If f is not continuous, it may not be possible to find tags for which $R(f,P) = U(f,P)$. Show however, that given an arbitrary $\epsilon > 0$ it is possible to pick tags for P so that

$$U(f,P) - R(f,P) < \epsilon \tag{22}$$

Solution. Assume f is a bounded function $f : [a,b] \to \mathbb{R}$.

a) If f is continuous, then each closed interval $[x_{k-1}, x_k]$ of the partition P will contain the supremum of $f(x)$ for x in that interval by the extremal value theorem. So we can simply set the tags $\{c_k\}$ for $R(f,P)$ to be the same as $\{M_k\}$ for $U(f,P)$. Then we will have $R(f,P) = U(f,P)$.

b) If f is not continuous, we may not have the supremum of $f(x)$ for x contained in some $[x_{k-1}, x_k]$ of the partition. However, since $M_k = \sup\{f(x) : x \in [x_{k-1}, x_k]\}$, then for all $\epsilon > 0$, we know there exists some $c_k \in [x_{k-1}, x_k]$ satisfying $M_k - c_k < \epsilon$. So for any $\epsilon > 0$, we are able to choose the collection of tags $\{c_k\}$ for which $U(f,P) - R(f,P) < \epsilon$.

■

Exercise 8.1.5. Complete the proof of Theorem 3.1

Proof. ($\Leftarrow$) Assume f is a bounded function $f : [a,b] \to \mathbb{R}$ with the property that for any $\epsilon > 0$, there exists a $\delta > 0$ such that, for any tagged partition $(P, \{c_k\})$ that is δ-fine, it follows that

$$|R(f,P) - A| < \epsilon. \tag{23}$$

We also know that for any $\epsilon > 0$ we have

$$U(f,P) - R(f,P) < \epsilon \tag{24}$$

and

$$|L(f,P) - R(f,P)| < \epsilon. \tag{25}$$

To show that $\int_a^b = A$ we need that

$$U(f,P) = L(f,P) = A. \tag{26}$$

As $U(f,P)$ and $L(f,P)$ are both within an ϵ distance from $R(f,P)$ for any $\epsilon > 0$, and since we can also choose a partition with small enough delta so that $R(f,P)$ is within an ϵ distance of A; it follows that there is only an ϵ distance separation between $U(f,P)$, $L(f,P)$ and A. Thus, we have $U(f,P) = L(f,P) = A$. □

Exercise 8.1.11. Show

$$F(b) - F(a) = \sum_{k=1}^{n} [F(x_k) - F(x_{k-1})]$$

Proof. Let $F : [a,b] \to \mathbb{R}$, and define a partition $P = [x_0, x_1, ..., x_n]$ of $[a,b]$. Assume without loss of generality that $x_0 = a$ and $x_n = b$ so we will let $F(b) - F(a) = F(x_n) - F(x_0)$. To show

$$F(x_n) - F(x_0) = \sum_{k=1}^{n} [F(x_k) - F(x_{k-1})], \tag{27}$$

we proceed by induction on n. Let $n = 1$. Then

$$\sum_{k=1}^{n} [F(x_k) - F(x_{k-1})] = F(x_1) - F(x_0). \tag{28}$$

So assume $F(x_n) - F(x_0) = \sum_{k=1}^{n}[F(x_k) - F(x_{k-1})]$ holds and let $m = n + 1$.

$$F(x_m) - F(x_0) = \sum_{k=1}^{m}[F(x_k) - F(x_{k-1})] \tag{29}$$

$$= \sum_{k=1}^{n}[F(x_k) - F(x_{k-1})] + [F(x_m) - F(x_{m-1})] \tag{30}$$

$$= F(x_n) - F(x_0) + F(x_m) - F(x_{m-1}) \tag{31}$$

$$= F(x_m) - F(x_0) \tag{32}$$

Note that in (31) we have applied the induction hypothesis and in (32) we use the fact that $m = n + 1$ so we have $F(x_n) = F(x_{m-1})$. Thus, we have

$$\sum_{k=1}^{n}[F(x_k) - F(x_{k-1})] = F(x_n) - F(x_0) \tag{33}$$

for all $n \in \mathbb{N}$. □

4 Metric Spaces and the Baire Category Theorem

4.1 Metric space basics

Definition 4.1. *Given a set X, a function $d : X \times X \to \mathbb{R}$ is a metric on X if for all $x, y \in X$ we have*

1. *$d(x, y) \geq 0$ with $d(x, y) = 0$ if and only if $x = y$,*
2. *$d(x, y) = d(y, x)$, and*
3. *for all $z \in X$, $d(x, y) \leq d(x, z) + d(z, y)$*

Definition 4.2. *Let (X, d) be a metric space. A sequence $(x_n) \subseteq X$ converges to $x \in X$ if for all $\epsilon > 0$ there exists and $N \in \mathbb{N}$ such that $d(x_n, x) < \epsilon$ whenever $n \geq N$.*

Definition 4.3. *A metric space (X, d) is complete if every convergent sequence in X converges to an element of X.*

The natural and assumed metric used in analysis when working with $C[0, 1]$ the set of continuous functions on the closed interval $[0, 1]$ is given by

$$||f - g||_\infty = \sup\{|f(x) - g(x)| : x \in [0, 1]\}$$

We obtain the sup-norm metric by setting $g = 0$. Then we have

$$||f||_\infty = \sup\{|f(x)| : x \in [0, 1]\}$$

Definition 4.4. *Let (X, d_1) and (Y, d_2) be metric spaces. A function $f : X \to Y$ is continuous at $x \in X$ if for all $\epsilon > 0$ there exists a $\delta > 0$ such that $d_2(f(x, f(y)) < \epsilon$ whenever $d_1(x, y) < \delta$.*

4.2 Topology on Metric Spaces

Definition 4.5. *Given $\epsilon > 0$ and an element x in a metric space (X, d), the ϵ-neighbourhood of x is the set $V_\epsilon = \{y \in X : d(x, y) < \epsilon\}$.*

Definition 4.6. *A set $O \subset X$ is open if for every $x \in O$ we can find a neighbourhood $V_\epsilon \subseteq O$, A point x is a limit point of a set A if every $V_\epsilon(x)$ intersects A in some point other than x. A set C is closed if it contains its limit points.*

Exercise 8.2.8. Let (X, d) be a metric space.

a) Verify that a typical ϵ-neighbourhood is an open set. Is the set

$$C_\epsilon(x) = \{y \in X : d(x, y) \leq \epsilon\}$$

a closed set?

b) Show that a set $E \subseteq X$ is open if and only if its complement is closed.

Solution. a) Consider the ϵ-neighbourhood $V_\epsilon(x)$ for $x \in X$. For any $v \in V_\epsilon(x)$, take $V_\epsilon(x)$ as as the ϵ-neighbourhood containing v. Certainly a set is fully contained within itself, so indeed ϵ-neighbourhoods are open.

Consider $C_\epsilon(x)$ for some $x \in X$. We claim $C_\epsilon(x)$ contains all its limit points. Suppose we have a limit point y with $d(x, y) > \epsilon$. Then set $\epsilon' = \frac{d(x,y)-\epsilon}{2}$. But then $V_{\epsilon'}(y)$ is an ϵ-neighbourhood of y containing no points of $C_\epsilon(x)$, so y is not a limit point. As $C_\epsilon(x)$ contains all of its limit points it must be closed.

b) Suppose E is open. Then all points in E have some ϵ-neighbourhood entirely contained in E. Let x be a limit point of E^c. Then all ϵ-neighbourhoods of x contain some point $y \in E^c$ with $y \neq x$. So there is no ϵ-neighbourhood of x entirely contained in E. Thus $x \notin E$ and $x \in E^c$. As x was an arbitrary limit point of E^c it follows that E^c is closed.

Conversely, suppose that E^c is closed. Then all limit points of E^c are contained in E^c. Choose $x \in E$, $x \notin E^c$. It must be that x is not a limit point of E^c. Thus there exists an ϵ-neighbourhood around x containing no points of E^c. Then this ϵ-neighbourhood is contained entirely in E. As $x \in E$ was arbitrary and such an ϵ-neighbourhood may always be found, it must be that E is open.

■

4.3 Baire Category Theorem

Definition 4.7. *Given a subset E of a metric space (X, d), the closure $\overline{E}$ of E is the union of E together with its limit points. The interior E° of E are those points in E for which there exists an ϵ-neighbourhood entirely contained in E.*

The closure of E is the smallest closed set containing E. The interior of E is the largest open set contained in E. Thus E is open if and only if $E = E^\circ$ and E is closed if and only if $E = \overline{E}$.

Definition 4.8. *A set $A \subseteq X$ is dense in the metric space (X, d) if $\overline{A} = X$. A subset E of a metric space (X, d) is nowhere-dense in X if the interior of the closure $\overline{E}^\circ$ is empty.*

Theorem 4.1. *Let (X, d) be a complete metric space and let $\{O_n\}$ be a countable collection of dense, open subsets of X. Then $\cap_{n=1}^{\infty} O_n$ is not empty.*

Exercise 8.2.14. Proof idea:

a) Give the details for why we know there exists a point $x_2 \in V_{\epsilon_1}(x_1) \cap O_2$ and an $\epsilon_2 > 0$ satisfying $\epsilon_2 < \frac{\epsilon_1}{2}$ with $V_{\epsilon_2}(x_2) \subseteq O_2$ and $\overline{V_{\epsilon_2}(x_2)} \subseteq V_{\epsilon_1}(x_1)$.

b) Proceed along this line and use the completeness of (X, d) to produce a single point $x \in O_n$ for every $n \in \mathbb{N}$.

Proof. Pick $x_1 \in O_1$. Because O_1 is open, there exists an $\epsilon_1 > 0$ such that $V_{\epsilon_1}(x_1) \subseteq O_1$.

Since O_2 is dense, $V_{\epsilon_1}(x_1)$ and O_2 are disjoint only if $V_{\epsilon_1}(x_1)$ is exclusively limit points of O_2. But since $V_{\epsilon_1}(x_1)$ is open this cannot be the case, so we must have $V_{\epsilon_1}(x_1) \cap O_2 \neq \emptyset$.

We can now pick $x_2 \in V_{\epsilon_1}(x_1) \cap O_2$. And since both $V_{\epsilon_1}(x_1)$ and O_2 are open we can always find $\epsilon_2 > 0$ small enough to satisfy $\epsilon_2 < \frac{\epsilon_1}{2}$ and

$$\overline{V_{\epsilon_2}(x_2)} \subseteq V_{\epsilon_1}(x_1) \cap O_2 \tag{34}$$

Continue in this way. Since O_{n+1} is dense in X we have $O_{n+1} \cap V_{\epsilon_n}(x_n) \neq \emptyset$. So we pick $x_{n+1} \in O_{n+1} \cap V_{\epsilon_n}(x_n)$. And since O_{n+1} and $V_{\epsilon_n}(x_n)$ are both open we know there exists $\epsilon_{n+1} > 0$ with $\epsilon_{n+1} < \frac{\epsilon_1}{n+1}$ and

$$\overline{V_{\epsilon_{n+1}}(x_{n+1})} \subseteq V_{\epsilon_n}(x_n) \cap O_{n+1} \tag{35}$$

Now we have a sequence of points (x_n) in X where $(\epsilon_n) \to 0$ and for any $n, m \in \mathbb{N}$ we have $|x_n - x_m| < \epsilon_1$. So (x_n) is a cauchy sequence, and by X complete it must converge to an element in X. As we have each $x_n \in O_n$, it must be that the limit of (x_n) is in $\cap_{n=1}^{\infty} O_n$, so the intersection is nonempty. □

Theorem 4.2. *A complete metric space is not the union of a countable collection of nowhere-dense sets.*

Exercise 8.2.15. Complete the proof of the theorem

Proof. Let (X, d) be a complete metric space and let $\{O_n\}$ be a countable collection of dense, open subsets of X. Define $C_n = O_n^c$. We claim C_n is nowhere-dense. So observe,

$$\begin{aligned}
C_n \text{ is nowhere-dense in X} &\iff \overline{C_n}^{\circ} = \emptyset && (36)\\
&\iff (\overline{C_n}^{\circ})^c = X && (37)\\
&\iff \overline{\overline{C_n}^{c}} = X && (38)\\
&\iff \overline{((C_n)^c)^{\circ}} = X && (39)\\
&\quad = \overline{(O_n)^{\circ}} = X && (40)\\
&\iff (O_n)^{\circ} \text{ is dense in } X && (41)\\
&\iff O_n \text{ is dense in } X && (42)
\end{aligned}$$

Thus as all O_n are dense in X, all C_n are nowhere-dense in X. Now $\{C_n\}$ is a countable collection of closed nowhere-dense sets. From above, we know $\cap_{n=1}^{\infty} O_n \neq \emptyset$. Through applying De Morgan's laws,

$$\begin{aligned}
(\cap_{n=1}^{\infty} O_n)^c &\neq \emptyset^c && (43)\\
\cup_{n=1}^{\infty} O_n^c &\neq X && (44)\\
\cup_{n=1}^{\infty} C_n &\neq X && (45)
\end{aligned}$$

Thus X is not the union of a countable collection of nowhere-dense sets. □

5 Euler's Sum

5.1 Wallis's Product

We encounter two different representations of the function $\sin(x)$. First the Taylor series representation

$$\sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + ..., \tag{46}$$

and second the infinite product representation

$$\sin(x) = x(1 - \frac{x}{\pi})(1 + \frac{x}{\pi})(1 - \frac{x}{2\pi})(1 + \frac{x}{2\pi}) \cdot ... \tag{47}$$

We do not have enough tools to prove the infinite product formula for $\sin(x)$. We can work with a special case where $x = \frac{\pi}{2}$.

Exercise 8.3.1. Show that when $x = \frac{\pi}{2}$, equation (47) is equivalent to

$$\frac{\pi}{2} = \lim_{n \to \infty} \Big(\frac{2 \cdot 2}{1 \cdot 3}\Big)\Big(\frac{4 \cdot 4}{3 \cdot 5}\Big)\Big(\frac{6 \cdot 6}{5 \cdot 7}\Big) \cdots \Big(\frac{2n \cdot 2n}{(2n-1) \cdot (2n+1)}\Big)$$

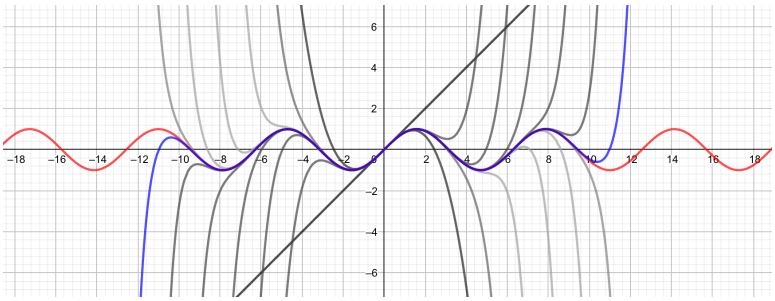


Figure 5: Taylor series representation of $\sin(x)$ out to $\frac{x^{25}}{25!}$

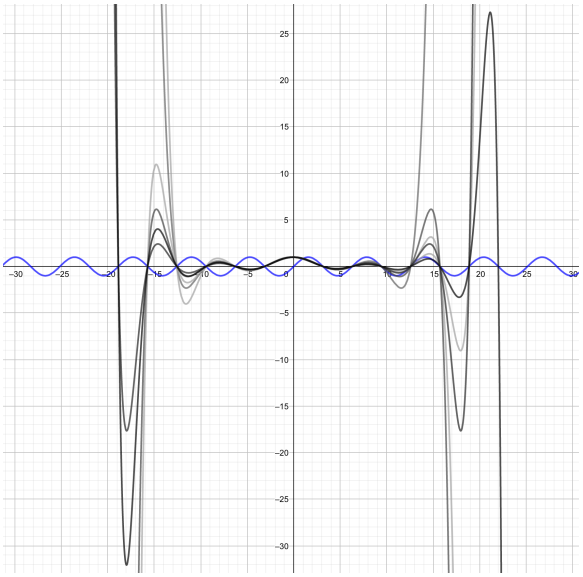


Figure 6: Infinite product representation of $\sin(x)$ out to $(1-\frac{x}{5\pi})(1+\frac{x}{5\pi})$

Solution. In equation (47), set $x = \frac{\pi}{2}$.

$$\sin\left(\frac{\pi}{2}\right) = \frac{\pi}{2}\left(1-\frac{\frac{\pi}{2}}{\pi}\right)\left(1+\frac{\frac{\pi}{2}}{\pi}\right)\left(1-\frac{\frac{\pi}{2}}{2\pi}\right)\left(1+\frac{\frac{\pi}{2}}{2\pi}\right)\cdot\ldots \tag{48}$$

$$1 = \frac{\pi}{2}\left(1-\frac{\pi}{2\pi}\right)\left(1+\frac{\pi}{2\pi}\right)\left(1-\frac{\pi}{4\pi}\right)\left(1+\frac{\pi}{4\pi}\right)\cdot\ldots \tag{49}$$

$$\Rightarrow \text{multiply through by } \frac{2}{\pi} \text{ and take the reciprocal (we know all terms are nonzero)} \tag{50}$$

$$\frac{\pi}{2} = \left(\frac{1}{1-\frac{\pi}{2\pi}}\right)\left(\frac{1}{1+\frac{\pi}{2\pi}}\right)\left(\frac{1}{1-\frac{\pi}{4\pi}}\right)\left(\frac{1}{1+\frac{\pi}{4\pi}}\right)\cdot\ldots \tag{51}$$

$$\frac{\pi}{2} = \left(\frac{1}{1/2}\right)\left(\frac{1}{3/2}\right)\left(\frac{1}{3/4}\right)\left(\frac{1}{5/4}\right)\cdot\ldots \tag{52}$$

$$\frac{\pi}{2} = \left(\frac{2}{1}\right)\left(\frac{2}{3}\right)\left(\frac{4}{3}\right)\left(\frac{4}{5}\right)\cdot\ldots \tag{53}$$

It follows then that

$$\frac{\pi}{2} = \lim_{n\to\infty}\left(\frac{2n\cdot 2n}{(2n-1)\cdot(2n+1)}\right) \tag{54}$$

■

We refer to this equation as Wallis's Product.

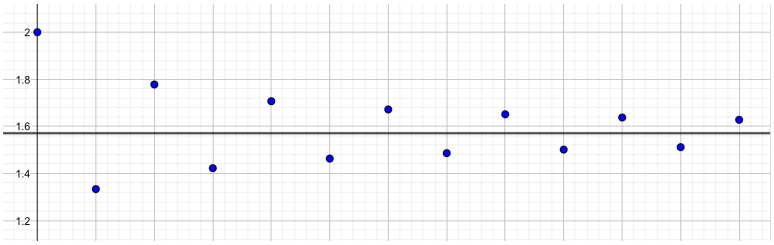


Figure 7: Wallis Product approximation of $\frac{\pi}{2}$ out to $n = 13$

Exercise 8.3.2. Assume $h(x)$ and $k(x)$ have continuous derivatives on $[a, b]$ and derive the integration by parts formula:

$$\int_a^b h(t)k'(t)dt = h(b)k(b) - h(a)k(a) - \int_a^b h'(t)k(t)dt$$

Proof. From the product formula for derivatives we may write

$$\frac{d}{dt}[h(t)d(t)] = \frac{d}{dt}[h(t)]k(t) + h(t)\frac{d}{dt}[k(t)] \tag{55}$$

Rearrange this equation to obtain,

$$h(t)\frac{d}{dt}[k(t)] = \frac{d}{dt}[h(t)d(t)] - \frac{d}{dt}[h(t)]k(t) \tag{56}$$

As $h(x)$ and $k(x)$ have continuous derivatives on $[a, b]$, we take the integral from a to b of both sides.

$$\int_a^b \big(h(t)\frac{d}{dt}[k(t)]\big)dt = \int_a^b \big(\frac{d}{dt}[h(t)d(t)]\big)dt - \int_a^b \big(\frac{d}{dt}[h(t)]k(t)\big)dt \tag{57}$$

Which simplifies to

$$\int_a^b h(t)k'(t)dt = [h(t)k(t)]_a^b - \int_a^b h'(t)k(t)dt \tag{58}$$

$$= h(b)k(b) - h(a)k(a) - \int_a^b h'(t)k(t)dt \tag{59}$$

and this is the formula we wanted. □

5.2 The Integral Form of the Remainder

Theorem 5.1. *Let f be differentiable $N + 1$ times on $(-R, R)$ and assume $f^{(N+1)}$ is continuous. Define $a_n = \frac{f^{(n)}(0)}{n!}$ for $n = 0, 1, ..., N$ and let*

$$S_N(x) = a_0 + a_1x + a_2x^2 + ... + a_Nx^N$$

For all $x \in (-R, R)$, the error function $E_N(x) = f(x) - S_N(x)$ satisfies

$$E_N(x) = \frac{1}{N!}\int_a^b f^{(N+1)}(t)(x - t)^N dt$$

Proof. The case $x = 0$ is easy to check, so take $x > 0$.

Exercise 8.3.9. Show

$$f(x) = f(0) + \int_0^x f'(t)dt$$

Then use a previous result (Ex 8.3.2) to show

$$f(x) = f(0) + f'(0)x + \int_0^x f''(t)(x-t)dt$$

Generalise so as to complete the proof.

First, we observe

$$\begin{aligned} f(x) &= f(0) + \int_0^x f'(t)d && (60)\\ &= f(0) + f(x) - f(0) && (61)\\ &= f(x) && (62) \end{aligned}$$

And now using integration by parts,

$$\begin{aligned} f(x) &= f(0) + f'(0)x + \int_0^x f''(t)(x-t)dt && (63)\\ &= f(0) + f'(0)x + [(x-t)f'(t)]_0^x - \int_0^x f'(t)\Big[\frac{d}{dt}(x-t)\Big]dt && (64)\\ &= f(0) + f'(0)x + [(x-t)f'(t)]_0^x + \int_0^x f'(t)dt && (65)\\ &= f(0) + \int_0^x f'(t)dt && (66) \end{aligned}$$

The equality of (66) was shown in (60).

We observe that (60) and (63) are beginning to look a lot like the $N = 0$ and $N = 1$ iterations for $f(x) = SN(x) + [\text{'the desired' } E_N(x)]$. But it appears we might be off by a factor of $N!$. We need one more iteration to really see the 'hidden' $\frac{1}{N!}$ that does not appear when we only look at $N = 0$ and $N = 1$!!(shock not factorial). So, let's take $N = 2$.

$$\begin{aligned} f(x) &= f(0) + f'(0)x + \frac{1}{2}f''(0)x^2 + \frac{1}{2}\int_0^x f'''(t)(x-t)^2dt && (67)\\ &= f(0) + f'(0)x + \frac{1}{2}f''(0)x^2 + \frac{1}{2}[(x-t)^2f''(t)]_0^x - \frac{1}{2}\int_0^x f''(t)\Big[\frac{d}{dt}(x-t)^2\Big]dt && (68)\\ &= f(0) + f'(0)x + \frac{1}{2}f''(0)x^2 + \frac{1}{2}[(x-t)^2f''(t)]_0^x - \frac{1}{2}\int_0^x f''(t)(-2(x-t))dt && (69)\\ &= f(0) + f'(0)x + \frac{1}{2}f''(0)x^2 - \frac{1}{2}f''(0)x^2 + \int_0^x f''(t)(x-t)dt && (70)\\ &= f(0) + f'(0)x + \int_0^x f''(t)(x-t)dt && (71) \end{aligned}$$

Now (71) is equivalent to (63).

Notice the factor of $N!$ is present even when it is a nontrivial value. Now we may generalise. For any $N = 0, 1, ...$ we have

$$f(x) = \left[\sum_{n=0}^{N} \frac{x^n}{n!}f^{(n)}(0)\right] + \frac{1}{N!}\int_0^x f^{(N+1)}(x-t)^N dt \quad (72)$$

Now, the summand term is simply $S_N(x)$. So we have

$$f(x) - S_N(x) = \frac{1}{N!}\int_0^x f^{(N+1)}(x-t)^N dt \quad (73)$$

Which shows that the error function $E_N(x) = f(x) - S_N(x)$ indeed has the desired property. □

6 Inventing the Factorial Function

The factorial function as we know it is defined only on $\mathbb{N}$, and possibly with the extension $0! = 1$. We wish to extend the factorial function in a meaningful way to all real numbers. That is we want a function $f(x) : \mathbb{R} \to \mathbb{N}$ with the property that $f(n) = n!$ for all $n \in \mathbb{N}$. So define f piecewise as

$$f(x) = \begin{cases} n! & \text{if } n \leq x < n+1, n \in \mathbb{N} \\ 1 & \text{if } x < 1 \end{cases} \tag{74}$$

This extension of $n!$ to $\mathbb{R}$ is missing the key property we would like of $n!$, namely that $n! = n(n-1)!$.

6.1 The Exponential Function

Definition 6.1. *Define*

$$E(x) = \sum_{n=0}^{\infty} \frac{x^n}{n!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + ... \tag{75}$$

We will show now that the function $E(x)$, based on a power series, will behave like e^x. First, we need to state some theorems related to convergence and differentiability of power series such as $E(x)$

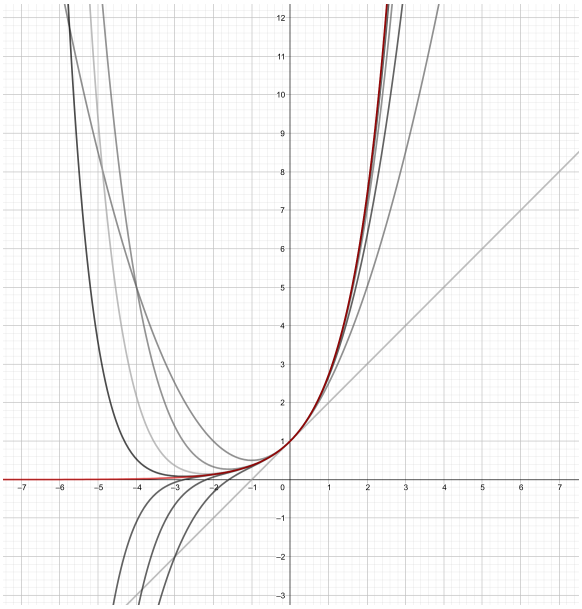


Figure 8: The exponential approximation of e^x out to $\frac{x^8}{8!}$

Lemma 6.1. *Given a series $\sum_{n=1}^{\infty} a_n$ with $a_n \neq 0$, if (a_n) satisfies*

$$\lim \left| \frac{a_{n+1}}{a_n} \right| = r < 1, \tag{76}$$

then the series converges absolutely.

Theorem 6.2. *Let $f_n \to f$ pointwise on the closed interval $[a, b]$, and assume that each f_n is differentiable. If(f_n') converges uniformly on $[a, b]$ to a function g, then the function f is differentiable and $f' = g$.*

Theorem 6.3. *If a power series $\sum)n = 1^{\infty} a_n x^n$ converges absolutely at a point x_0, the it converges uniformly on the closed interval $[-c, c]$ where $c = |x_0|$.*

Exercise 8.4.2. Verify that the series converges absolutely for all $x \in \mathbb{R}$, that $E(x)$ is differentiable on $\mathbb{R}$ and that $E'(x) = E(x)$.

Proof. For some fixed $x \in \mathbb{R}$, take $a_n(x) = \frac{x^n}{n!}$. Then $E(x) = \sum_{n=0}^{\infty} a_n(x)$. We may apply lemma 6.1 to see if $E(x)$ converges absolutely:

$$\lim_{n\to\infty} \left|\frac{a_{n+1}(x)}{a_n(x)}\right| = \lim_{n\to\infty} \left|\frac{\frac{x^{n+1}}{(n+1)!}}{\frac{x^n}{n!}}\right| \tag{77}$$

$$= \lim_{n\to\infty} \left|\frac{x^{n+1}n!}{x^n(n+1)!}\right| \tag{78}$$

$$= \lim_{n\to\infty} \left|\frac{x}{n+1}\right| \tag{79}$$

$$= 0 \tag{80}$$

As $\lim_{n\to\infty} \left|\frac{a_{n+1}(x)}{a_n(x)}\right| < 1$, we know $E(x)$ converges absolutely.

To verify that $E(x)$ is differentiable on $\mathbb{R}$. we will apply theorem 6.2. We first need that each $a_n(x)$ is differentiable:

$$a_n'(x) = \left[\frac{x^n}{n!}\right]' \tag{81}$$

$$= \frac{1}{n!}[x^n]' \tag{82}$$

$$= \frac{n}{n!}x^{n-1} \tag{83}$$

$$= \frac{x^{n-1}}{(n-1)!} \tag{84}$$

$$= a_{n-1}(x) \tag{85}$$

We now need to establish uniform convergence of $E(x)$. We will use theorem 6.3. Since $E(x)$ converges absolutely for all $x \in \mathbb{R}$, $E(x)$ for any $x_0 \in \mathbb{R}$, set $c = |x_0|$ and we will have that $E(x)$ converges uniformly on the interval $[-c, c]$. Thus $E(x)$ converges uniformly on all of $\mathbb{R}$. Now, since $E(x)$ converges uniformly and each $a_n(x)$ is differentiable, with $a_n'(x) = a_{n-1}(x)$, we know that $E(x)$ is differentiable and that $E'(x) = E(x)$. □

Lemma 6.4. *Let $\sum_{i=1}^{\infty} a_i$ and $\sum_{j=1}^{\infty} b_i$ be two infinite series that converge absolutely to A and B respectively. Then*

$$\sum_{i=1}^{\infty}\sum_{j=1}^{\infty} a_i b_j = \sum_{j=1}^{\infty}\sum_{i=1}^{\infty} a_i b_j = \sum_{k=2}^{\infty} d_k = AB \tag{86}$$

where $d_k = a_1 b_{k-1} + a_2 b_{k-2} + \cdots + a_{k-1} b_1$.

Exercise 8.4.3. Show the following for all $x, y \in \mathbb{R}$

1. $E(x+y) = E(x)E(y)$
2. $E(0) = 1$
3. $E(-x) = \frac{1}{E(x)}$
4. $E(x) > 0$

Proof. Given $E(x) \sum_{n=0}^{\infty} \frac{x^n}{n!}$, we have

$$E(x+y) = \sum_{n=0}^{\infty} \frac{(x+y)^n}{n!} \tag{87}$$

$$= \sum_{n=0}^{\infty} \frac{1}{n!} \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k} \tag{88}$$

$$= \sum_{n=0}^{\infty}\sum_{k=0}^{n} \frac{x^k y^{n-k}}{k!}(n-k)! \tag{89}$$

By lemma 6.4 we have that since $\sum_{k=0}^{\infty} \frac{x^k}{k!} = E(x)$ and $\sum_{k=0}^{\infty} \frac{y^k}{k!} = E(y)$, we have that

$$E(x+y) = \sum_{n=0}^{\infty} \sum_{k=0}^{n} \frac{x^k y^{n-k}}{k!(n-k)!} = E(x)E(y) \tag{90}$$

.

We can now use this fact to show that $E(0) = 1$. Take $x = 0, y \neq 0$. Then $E(y) = E(0+y) = E(0)E(y)$. So we must have $E(0) = 1$.

This will help us show $E(x)^{-1} = E(-x)$:

$$1 = E(0) = E(x-x) = E(x)E(-x). \tag{91}$$

Thus we must have $E(x)^{-1} = E(-x)$.

Finally, if $x \geq 0$ we have $E(x) \geq 1 > 0$. Otherwise, if $x < 0$, we have $E(x) = E(-x)^{-1} > 0$. Thus $E(x) > 0$ for all $x \in \mathbb{R}$. □

Exercise 8.4.4. Define $e = E(1)$. Show that $E(n) = e^n$ and $E(\frac{m}{n}) = (\sqrt[n]{e})^m$ for all $m, n \in \mathbb{Z}$.

Proof. Since $e = E(1)$ and $E(x+y) = E(x)E(y)$, we have

$$E(n) = E(\underbrace{1+1\cdots+1}_{n}) = \underbrace{E(1)E(1)\cdots E(1)}_{n} = e^n \tag{92}$$

. Similarly,

$$E\Big(\frac{m}{n}\Big) = E\Big(\underbrace{\frac{1}{n}+\frac{1}{n}\cdots+\frac{1}{n}}_{m}\Big) = \underbrace{E\Big(\frac{1}{n}\Big)E\Big(\frac{1}{n}\Big)\cdots E\Big(\frac{1}{n}\Big)}_{m} = e^{m/n}. \tag{93}$$

□

6.2 The Functional Equation

Exercise 8.4.8. Inspired by the fact that $0! = 1$ and $1! = 1$, let $h(x)$ satisfy

1. $h(x) = 1$ for all $0 \leq x \leq 1$
2. $h(x) = x \cdot h(x-1)$ for all $x \in \mathbb{R}$.

a) Find a formula for $h(x)$ on $[1,2]$, $[2,3]$ and $[n, n+1]$ for arbitrary $n \in \mathbb{N}$.

b) Do the same for $[-1,0]$, $[-2,-1]$ and $[-n, -n+1]$.

c) Sketch h over the domain $[-4,4]$.

Solution. a) With the given conditions on $h(x)$ we begin with

$$h(x) = \begin{cases} 1 & \text{if } x \in [0,1] \\ x & \text{if } x \in [1,2] \\ x(x-1) & \text{if } x \in [2,3] \\ x(x-1)(x-2) & \text{if } x \in [3,4] \end{cases} \tag{94}$$

This allows us to cover the arbitrary case of positive real numbers. For $x \in [n, n+1]$, where $n \in \mathbb{N}$,

$$h(x) = \prod_{i=0}^{n-1} (x-i) \tag{95}$$

b) For negative real numbers, we take

$$h(x) = \begin{cases} \frac{1}{x+1} & \text{if } x \in [-1, 0] \\ \frac{1}{(x+1)(x+2)} & \text{if } x \in [-2, -1] \\ \frac{1}{(x+1)(x+2)(x+3)} & \text{if } x \in [-3 - 2] \end{cases} \tag{96}$$

Now the values of $-1, -2, -3$, etc. do not exist. But the desired property $h(x) = x \cdot h(x-1)$ does hold. So we extend this definition to the arbitrary case of negative real numbers. For $x \in [-n, -n+1]$, where $n \in \mathbb{N}$,

$$h(x) = \prod_{i=1}^{n} \frac{1}{x+i} \tag{97}$$

c) The piece-wise graph of $h(x)$ on $[-4, 4]$ shown here.

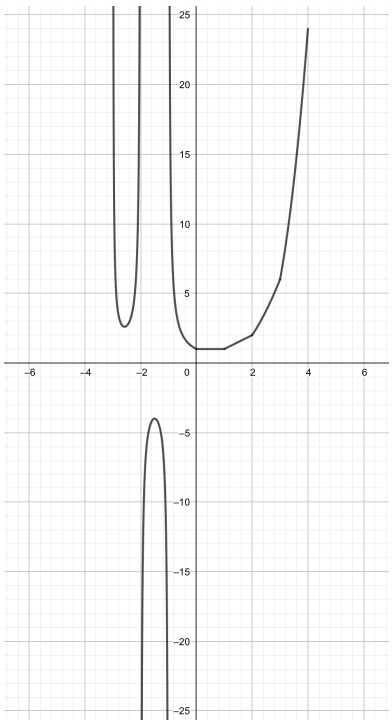


Figure 9: First approximation of $h(x)$ on $[-4, 4]$

■

We observe that $h(x) = x!$ for natural numbers x. But $h(x)$ is not continuous for $x < 0$ and is non-differentiable at several points even for $x > 0$.

6.3 Constructing the Gamma Function

Definition 6.2. *Let $f(x,t)$ be a function of two variables defined for all $a \leq x \leq b$ and $c \leq t \leq d$. Then the domain of f is a rectangle $D \subset \mathbb{R}^2$. In $\mathbb{R}^2$ we will use the standard extension of the distance formula from $\mathbb{R}$. Define the distance between two points (x_0, y_0) and (x_1, y_1) as $||(x_0, y_0) - (x_1, y_1)|| = \sqrt{(x_1 - x_0)^2 + (t_1 - t_0)^2}$*

Definition 6.3. *A function $f : D \to \mathbb{R}$ is continuous at (x_0, t_0) if for all $\epsilon > 0$ there exists $\delta > 0$ such that whenever $||(x,t) - (x_0, t_0)|| < \delta$ it follows that $|f(x,t) - f(x_0, t_0)| < \epsilon$.*

Theorem 6.5. *If f is continuous on $[a, b]$, then f is integrable on $[a, b]$.*

Exercise 8.4.12. Assume the function $f(x,t)$ is continuous on $D = \{(x,t) : a \leq x \leq b, c \leq t \leq d\}$. Explain why the function

$$F(x) = \int_c^d f(x,t)dt$$

is properly defined for all $x \in [a,b]$.

Proof. Suppose $f(x,t)$ is continuous on $D = \{(x,t) : a \leq x \leq b, c \leq t \leq d\}$. Then for any $(x_0,t_0) \in D$ and any $\epsilon > 0$ there exists $\delta > 0$ such that whenever $||(x,t) - (x_0,t_0)|| < \delta$ we have that $|f(x,t) - f(x_0,t_0)| < \epsilon$. Fix $x = x_0$. Then $|t - t_0| = ||(x,t) - (x_0,t_0)|| < \delta$ so we still have $|f(x,t) - f(x_0,t_0)| < \epsilon$. Hence $f(x_0,t)$ is a continuous function of t for $t \in [c,d]$ and thus by theorem 6.5, f is integrable on $[c,d]$. As we chose x_0 arbitrarily, it is true that $F(x) = \int_c^d f(x,t)dt$ is well defined for all $x \in [a,b]$. □

Definition 6.4. *For $x \geq 0$, define the factorial function*

$$x! = \int_0^\infty t^x e^{-t} dt \tag{98}$$

This gives us the well known gamma function

$$\Gamma(x) = (x-1)! = \int_0^\infty t^{x-1} e^{-t} dt \tag{99}$$

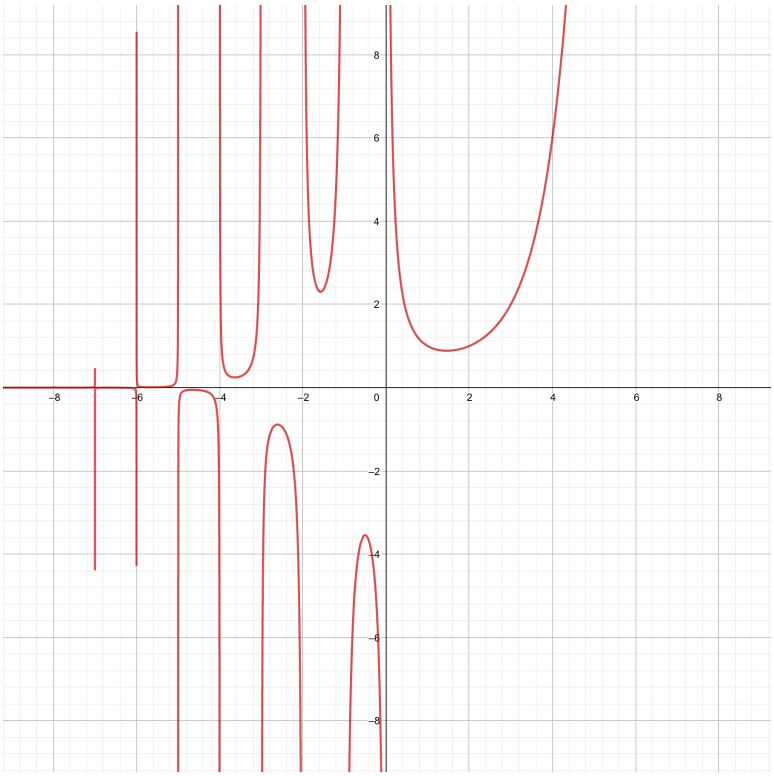


Figure 10: The gamma function

Exercise 8.4.20. a) Show that $x!$ is an infinitely differentiable function on $(0,\infty)$ and produce a formula for the n^{th} derivative.

b) Use the integration by. parts formula to show $x!$ satisfies $(x+1)! = (x+1)x!$.

Solution. a) Given

$$x! = \int_0^\infty t^x e^{-t} dt, \tag{100}$$

we have

$$[x!]' = \frac{d}{dx}\left[\int_0^\infty t^x e^{-t} dt\right] \tag{101}$$

$$= \int_0^\infty t^x e^{-t} \ln(t) dt \tag{102}$$

$$[x!]'' = \frac{d}{dx}\left[\int_0^\infty t^x e^{-t} \ln(t) dt\right] \tag{103}$$

$$= \int_0^\infty t^x e^{-t} [\ln(t)]^2 dt \tag{104}$$

We can generalise now for an arbitrary n^{th} derivative:

$$[x!]^{\{n'\}} = \int_0^\infty t^x e^{-t} [\ln(t)]^n dt \tag{105}$$

b) Again, given

$$x! = \int_0^\infty t^x e^{-t} dt, \tag{106}$$

we have

$$(x+1)! = \int_0^\infty t^{x+1} e^{-t} dt \tag{107}$$

Let $u = t^{x+1}$ and $v = -e^{-t}$. Then applying the integration by parts formula:

$$\int_a^b u dv = [uv]_a^b - \int_a^b v du \tag{108}$$

gives us

$$(x+1)! = \left[-e^{-t} t^{x+1} \right]_{t=0}^\infty - \int_0^\infty (x+1)(t^x)(-e^{-t}) dt \tag{109}$$

Since $\lim_{t\to\infty} \left[-e^{-t} t^{x+1} \right] = 0$, we have

$$(x+1)! = (x+1) \int_0^\infty t^x e^{-t} dt = (x+1)x! \tag{110}$$

■

7 Conclusion

The topics presented in this paper may be a somewhat random selection, but I do think the concepts are connected in that they are surface level intuitive but deeply complex. Many people think of mathematics as occurring perfectly naturally, but seeing the work that goes into creating e^x and $x!$ makes me think otherwise. The most inspiring part of this project was extending the factorial. This is an idea I remember considering while studying discrete mathematics, and assuming it could not be done or at least would not be meaningful. Something I still do not quite understand is how we came up with the gamma function. I definitely believe it is the extension of the factorial we were looking for, but who thought to use the exponential? There is still more detail hidden beneath what I have presented here and what Abbott and Rudin presented to me.

References

[1] Stephen Abbott, *Understanding Analysis*, second edition ed., Undergraduate Texts in Mathematics, no. 666, Springer, 2016.

[2] David M. Bressoud, *Wrestling with the Fundamental Theorem of Calculus*, (2003).

[3] Walter Rudin, *Principle of Mathematical Analysis*, International series in pure and applied mathematics, McGraw-Hill Inc., 1976.

Challenges in Mathematical Formalization

Sarah Dennis

December 10, 2017

Abstract

The nineteenth century was a period of mathematical revolution. Developments in set theory led Georg Cantor, and subsequently Bertrand Russell, to discover the paradoxical nature of self-referential statements and infinite sets. In the subsequent period of uncertainty in mathematics, David Hilbert proposed demonstrating the proficiency of number theory; in particular, proving syntactically the completeness and consistency of the Peano Natural Numbers. Only thirty years later, Kurt Gödel published his First Incompleteness Theorem, which would show that the if the natural numbers are consistent, then they are incomplete, and consequently there exist true statements in this system that are unprovable within the system. Gödel followed up this work with his Second Incompleteness Theorem, demonstrating that in fact no formal system can prove its own consistency. Gödel's work surely shook the mathematical community, but Hilbert's *non ignorabimus* ideology is still seen today in the continued prevalence of computer proof assistants. Proof assistants allow us to formalize our proofs so as to be sure they are consistent with the axioms of some specified formal system. However, by this very nature, computer proof assistants have the same incompleteness as any of our formal arithmetic systems, in that there are true statements that will be unprovable. And so, while it may seem that computer formalization is taking over the task of mathematician, computers remain unable to analyze a statement's truth value from outside of a system as the human brain is able to.

1 Introduction

Mathematics is so often depicted as a discipline in which every possible question has a single correct answer. In my experience, if you ask a child why they enjoy mathematics, they will comment on the objectivity of math and the reliability of its problems to have some absolute truth that you as a student will be able to uncover. Then I have found, students of mathematics reach a point at which we begin to ask, "But *why* is this the right answer?". I know that when I first asked this question, I received the response "At some point a decision regarding the axioms of mathematics had to be made, and this was the norm we chose". I found this answer dreadfully unsatisfying. But of course mathematics cannot exist as some choose-your-own-adventure novel. We require universally agreed upon axioms in order to derive theorems that (assuming their basis in logic is agreeably sound) will be unilaterally accepted. Such theorems allow us to develop the complexity of our mathematical understanding. But how do we know that these axioms were chosen correctly, that we made the right decision? And how

can we be sure our system has all the axioms to needed to answer our complex questions? These were questions David Hilbert was strongly invested in answering at the start of the twentieth century.

2 Crisis in mathematics

2.1 A period of rapid advancement

The nineteenth century is often regarded as the second birth of mathematics. Our knowledge and understanding grew rapidly and broadly, and consequently the sub-disciplines of mathematics were pulled closer and closer, becoming more united and overlapping than ever. The period saw developments in geometry beyond Euclid's axioms, explorations into complex analysis, and elliptic and differential equations, increasing trends towards formalization in number theory, and a push towards investigating the infinite particularly in set theory. In general, there is a strong shift towards abstraction, a de-emphasis on calculation, and a growing confidence in working with the infinite across disciplines.[1] And herein lies the crisis of the nineteenth century in mathematics: with these broad and staggering advances in the mathematical sphere

> "a climate of opinion was generated in which it was tactically assumed that each sector of mathematical thought can be supplied with a set of axioms sufficient for developing systematically the endless totality of true propositions about the given area of inquiry." [2]

Each success strengthened the air of optimism spurring further innovation, but the mathematical systems in place were beginning to strain under the pressure of all of this new knowledge. As our depth of understanding grew, it became unclear as to whether the formal systems in place were sufficient to support the new discoveries and constructions.

In 1899, Georg Cantor discovers inconsistencies in his set theory when we consider infinite sets. First, in the notion of a power set, it is clear upon consideration that for any set A, the set of all subsets of A (i.e. the power set of A) has a strictly greater cardinality than that of A itself. Consequently, Cantor proves that the cardinality of the power set of natural numbers is equal to the cardinality of the real numbers. And furthermore, that if S is any set, then S cannot contain elements of all cardinalities – that in fact, there is a strict upper bound on the cardinalities of the elements of S [3]. At the time of Cantor's work, many mathematicians still shied away from the concept of infinities, and so

[1] Drawn from discussion found in Avigad & Reck, 2001
[2] Nagel & Newman, 2001, p. 24
[3] Farlow, 2008

Cantor's discovery was severely downplayed. At least David Hilbert would recognize the significance of Cantor's work. Hilbert began to push for the complete formalization of the natural numbers in the hope that, should more inconsistencies arise in set theory, a complete and provably consistent formal system could take its place.

2.2 David Hilbert's Program

> "Take any definite unsolved problem ... However unapproachable these problems may seem to us, and however helpless we stand before them, we have, nevertheless, the firm conviction that their solution must follow by a finite number of logical processes ... We hear within us the perpetual call: There is the problem. Seek its solution. You can find it by pure reason, for in reason there is no ignorabimus." - David Hilbert[4]

In 1900, David Hilbert is thirty years old and is recognized as one of the foremost mathematicians of his day.[1] Hilbert is famous for his notion of *no ignorabimus*. In direct response to the *ignoramus et ignorabimus* movement, proclaiming "We do not know and we will not know". Instead, Hilbert proposes *no ignorabimus*, pronouncing "Wir müssen wissen — wir werden wissen" ("We must know — we will know"). This determination and optimism gives us insight into Hilbert's reaction to Cantor's paradox in set theory. The horror of an inconsistent system could not disrupt Hilbert's *no ignorabimus* again; he was determined to secure the reliability of the Peano Axioms of the natural numbers (found in section 4.1). Hilbert was committed to proving the natural numbers to be a formal system in which we could, and would, know everything.

In a lecture at the Second International Congress of Mathematicians in Paris, Hilbert released his list of twenty three problems for the mathematical community, centered around the formalization of mathematics. Tasks in Hilbert's plan included:

- Formalize the natural numbers; construct a precise formal language that can be manipulated according to a specific and finite set of axioms via certain rules of inference.
- Prove the completeness of the system of natural numbers; find a proof that all true statements one can formulate in the language of the system can be proved using only the axioms and rules of inference of the system.
- Prove the consistency of the natural numbers; find a proof that no contradiction can be derived from the axioms of the system. Such a proof ought to be via finite reasoning, and in the formal language of the system.

[4]Franzen, 2005, p. 16

- Prove that the natural numbers are decidable; find an algorithm for deciding the truth or falsity of any formal statement in the system.

(Formal definitions of several terms referenced above in Hilbert's plan can be found in section 3.3).

2.3 Russell's Paradox

Only a year on from Hilbert's announcement, logician Bertrand Russell discovers an extension of Cantor's inconsistency in infinite set theory to finite set theory, developing the now well know paradox.

> **Russell's Paradox:** Call the set of all the sets that do not belong to themselves R. Does R belong to itself? By the definition of R, a set belongs to it if, and only if, it does not belong to itself. If we apply this rule to R itself, we obtain: R belongs to itself only if it does not belong to itself. This is a contradiction: a proposition and its negation cannot both be true at the same time.[5]

Where Cantor's paradox was hidden in the obscure nature of infinities, Russell's paradox thoroughly shook the mathematical community, particularly those in set theoretic study, with its applicability to the simplest of sets. For example,

> A barber who lives in a small village vows to give a haircut to precisely those villagers who do not cut their own hair. But now this barber is in a quandary: must she cut her own hair, or not? If she gives herself a haircut, then, according to the vow she took, she cannot cut her own hair. But, if she will not give herself a haircut, then she must do so![5]

This colloquial demonstration of Russell's paradox had mathematicians questioning why they had not thought of such problematic situations as these before.

> Behind Russell's paradox, as behind the original paradox of Cantor, is an assumption called the 'Axiom of comprehension'. This states that every property defines a set ... This assumption, however, leads to circular definitions. As Russell's paradox shows, with the help of the axiom of comprehension, we can define a set for which the relationship 'belonging to itself' is self-defined.[6]

How would the mathematical community react to this contradiction in elementary set theory?

With each paradox uncovered, and with each being "constructed by means of familiar and seemingly cogent modes of reasoning, mathematicians came to realize that in developing consistent systems,

[5] Aharoni, 2015, p. 206
[6] Aharoni, 2015, p. 207

familiarity and intuitive clarity are weak reeds to lean on."[7] There must be some way to prove that the intuition underlying our choice of axioms is sound, and that we are not wasting time proving theorems in an inconsistent system. Hilbert's desire for a completeness proof of the natural numbers was becoming of more ubiquitous interest. And yet,

> in September 1930 a conference on the foundations of mathematics was held in Konigsberg, and was attended by some of the best mathematicians of Europe. An announcement given at the end by a young, shy and slightly-built mathematician went hardly noticed.[8]

Kurt Gödel would shatter Hilbert's program in every way possible.

3 Gödel's Incompleteness Theorems

3.1 The life of Kurt Gödel

Kurt Gödel born in 1906 in Moravia (now the Czech Republic) attended the University of Vienna where, emigrating to the United States in 1940. Nine years prior to his relocation, Gödel published his first incompleteness theorem sending "shock waves through logic and the philosophy of mathematics."[9] Hilbert's plan was widely viewed as achievable, "it was generally believed that such a program was in principle possible, until the incompleteness theorem destroyed that hope."[9] Gödel's work regarding completeness and computability was continued by Alan Turing; the development of the Turing machine, being physical proof that an algorithm to enumerate all proofs of a formal system could not exist, would be the final blow Hilbert's plan.

Gödel's further work in mathematics include his demonstration in 1938 that the axiom of choice and the general continuum hypotheses are consistent with the Zermelo-Frankel axioms for natural numbers. Perhaps most impressive of all, is Gödel's staggering ability to work in realms of science and mathematics beyond that of pure logic, where he is arguably the most successful mathematician of the twentieth century. While at the University of Vienna, Gödel befriended Albert Einstein, and would later go on to establish "the existence of models of Einstein's Field Equations that permit time travel into the past".[9]

[7] Nagel & Newman, 2001, p. 24
[8] Aharoni, 2015, p. 213
[9] "Part VI Mathematicians," 2008, p. 819

3.2 Two incompleteness theorems

Theorem 1. (First Incompleteness Theorem)[10] *Any consistent formal system S within which a certain amount of elementary arithmetic can be carried out is incomplete with regard to statements of elementary arithmetic: there are such statements in S which can neither be proved, nor disproved in S.*

Theorem 2. (Second Incompleteness Theorem)[11] *For any consistent formal system S within which a certain amount of elementary arithmetic can be carried out, the consistency of S cannot be proved in S itself.*

3.3 The Language of Incompleteness

Relevant terminology from both incompleteness theorems and from Hilbert's plan includes [12]:

I. *A formal system:* A formal system is a system of axioms (expressed in some formally defined language) and of rules of reasoning (also called inference rules) used to derive the theorems of the system.

II. *Consistent:* In a consistent formal system, any statement must be either true or false, not both. If a statement is true, it becomes a theorem of the system. Furthermore, if both a statement and its negation are both derivable, the formal system is inconsistent.

III. *Demonstrable, derivable, deducible:* A statement is demonstrable if there exists a formal proof of its truth value within the system in which it is stated.

IV. *Complete:* A formal system is complete if all true statements in the system are deducible from a finite set of axioms via the rules of inference. A formal system is incomplete if there exists a true statement expressible in the system that is not provable within the system.

V. *Decidable:* A statement is decidable if it has a defined truth value found via proof in the system. In an inconsistent formal system, every statement is decidable.

VI. *A certain amount of elementary arithmetic:* The formal systems Gödel refers to are those with at least enough arithmetic language so as to be able to mirror the arithmetic statements of our common languages.

[10] Franzen, 2005, p. 16
[11] Franzen, 2005, p. 34
[12] Franzen, 2005, p. 17 - 24

3.4 A Toolbox for Gödel's Proofs

3.4.1 Gödel Numbering and PM

For the purpose of proving his incompleteness theorems Gödel describes a formalized calculus, which we shall refer to as PM.[13] In this system, all customary arithmetical notations can be expressed and familiar arithmetical relationships can be established. We will often refer throughout the next section and in Gödel's proofs to *metamathematical statements*, which are simply statements expressing mathematical ideas in a common language. PM is constructed such that we can translate or encompass any meta-mathematical statement into the formulaic language of PM.

With the formal system PM in place, Gödel introduces his system of Gödel numbering.[13] In PM, and in any formal system, it is possible to assign a unique integer to each symbol, statement (being a finite sequence of symbols), and proof (being a finite sequence of statements) in the language of PM. Despite the fact that there exist infinitely many unique statements in PM, each has a definite length. And so the set of all symbols, statements and proofs in PM is enumerable and may be ordered by length; hence this set is countable and can be mapped onto the positive integers. The integer assigned to each a symbol, formula, or proof is called its Gödel number. Gödel numbering is thoroughly arbitrary; the specific number assigned to a formula is irrelevant considering our desire to generalize Gödel's proof to varying formulas of PM. It is simply that we have the ability to determine a Gödel number for any formula of PM that will be key to the argument behind Gödel's incompleteness theorems.

3.4.2 The dem operation

Gödel demonstrates that, as every expression in PM is uniquely associated with a particular Gödel number, we are able to construct meta-mathematical statements describing specific formulas in PM by referring to each formula's corresponding Gödel number. This allows us to translate typographical relationships between formulas in PM, expressed in the language of PM, into arithmetic relationships between the Gödel numbers of formulas in PM, expressed in metamathematical language.[14]

For example, consider the metamathematical statement: *The sequence of formulas with Gödel number x is a proof in PM of the formula with Gödel number z.* This statement expressing the

[13] PN and Gödel numbering drawn from Nagel & Newman, 2001, ch VII part A
[14] dem and sub operations drawn from Nagel & Newman, 2001, ch VII part B

relationship between x and z we can formalize in PM as

$$\text{dem } (x, z). \tag{1}$$

The existence of formula (1) inside PM is crucial to Gödel's incompleteness theorems as it shows that true metamathematical assertions of the form *such-and-such demonstrates so-and-so by the rules of PM* are faithfully reflected within the language of PM. By the same nature, we are able to express statements of the form *such-and-such does not demonstrate so-and-so by the rules of PM*, in the language of PM:

$$\sim \text{dem } (x, z). \tag{2}$$

The dem operation is major building block in constructing a problematic self-referential statement in PM.

3.4.3 The sub operation

One final important tool used in Gödel's proof is the idea of substituting a string's own Gödel number into the string itself, and then taking the Gödel number of the resultant formula.[14] The new Gödel number can be expressed in terms of the old Gödel number in the language of PM as:

$$\text{sub } (x, *y, x), \tag{3}$$

where the notation $*y$, refers to the Gödel numerical value of the variable y, rather than the variable itself. Formula 3 represents the Gödel number of the formula obtained by taking the formula x with Gödel number $*x$ and, wherever there are occurrences of the variable y in the formula x, replacing them by the Gödel number $(*x)$ of x. That is to say, some statement with Gödel number $*x$ contains the variable y, and we are able to replace y with the statement x's Gödel number. This statement is no longer exactly the statement x, but it *is* still a valid statement in PM, and as such has its own unique Gödel number; formula (3) serves as a representation of this Gödel number in PM.

3.5 Proving Gödel's First Incompleteness Theorem

There are four primary steps in Gödel's proof of the first incompleteness theorem.[15] First, Gödel outlines how to construct the metamathematical statement *The formula G is not demonstrable using the rules of PM*; where the formula G states *The formula that has Gödel number *g is not demonstrable.* Gödel also shows that G is demonstrable if, and only if, $\neg G$ is demonstrable. But, as we well know, if both a formula and its negation are formally demonstrable in a system, then that system is incomplete. In other words, if PM is consistent, then G is formally undecidable. Then, Gödel will show that while G is not demonstrable, it is in fact true. Finally, it is highlighted that since G is both true and formally undecidable in PM, PM must be incomplete.

3.5.1 Step 1.

→ Construct a formula G of PM that represents the meta-mathematical statement: *The formula G is not demonstrable using the rules of PM.*[16]

By prefixing formula (1) with the existential quantifier,we obtain

$$\exists x \text{ dem } (x, z), \tag{4}$$

a formula of PM that states *"There exists a sequence of formulas with Gödel number *x that constitutes a proof of the formula with Gödel number *z"*. In other words, we now have a formalization of the statement *"The formula with Gödel number *z is demonstrable"*. By negating (4) we obtain,

$$\neg\exists x \text{ dem } (x, z), \tag{5}$$

which now conveys *"The formula with Gödel number *z is not demonstrable"*. This formula is still too vague to embody the meta-mathematical statement we set out to construct; the formula with Gödel number $*z$ could be any formula, whereas we particularly want to show that the formula G is not demonstrable.

To this end, consider the formula

$$\text{sub } (y, *y, y). \tag{h}$$

[15] Outline of Gödel's proof from Nagel & Newman, 2001, ch VII part C
[16] Step 1 from Nagel & Newman, 2001, ch VII part B, p. 95 - 98

We defined the sub operation in (3), and by its construction, h represents the formula resulting from the manipulation of the formula y such that all occurrences of y's Gödel number $*y$ in y are replaced by the formula y itself. Now, let z in (5) be this specific formula h of PM, and by making this substitution we obtain the formula,

$$\neg\exists x \text{ dem } (x,\ \text{sub } (y, *y, y)). \tag{n}$$

This eliminates the ambiguity of the variable z, and now we have a formula of PM stating, *"There does not exist a proof with Gödel number *x of the formula with Gödel number* $*h = \text{sub } (y, *y, y)$". In other words, *"The formula with Gödel number *h is not demonstrable"*. This brings us one step closer to our initial goal since we now know specifically what the Gödel number *h refers to.

The formula n is still not yet well defined however, as it contains the variable y whose value we have not specified. As n is a formula in PM, it has a unique Gödel number which we shall call $*n$. Now, by replacing all occurrences of the variable y in n with the numerical value $*n$ we create the new formula,

$$\neg\exists x \text{ dem } (x,\ \text{sub } (*n, *y, *n)). \tag{G}$$

The meaning of G is definite since there are now no unquantified variables remaining. However, is G the certain self-referential statement we are looking for.

The statement G also occurs in PM, and so G has a Gödel number we will refer to as $*g$. Recall that formula

$$\text{sub } (*n, *y, *n). \tag{6}$$

has the Gödel number of whichever formula that results when we substitute $*n$ for the variable with Gödel number $*y$ inside the formula whose Gödel number is $*n$. But G was obtained in exactly this manner; we began with formula n, having Gödel number $*n$, and replaced all occurrences of y in n with the Gödel number $*n$ and called this equation G. And so (6) is in fact the Gödel number $*g$ for G. Finally, the formula G then states *"There does not exist a proof with Gödel number *x of the statement with Gödel number *g"*, or in other words, *"The formula with Gödel number *g is not demonstrable"*; this is exactly the statement we wished to construct.

3.5.2 Step 2.

$\rightarrow$ Show that G is demonstrable if, and only if, its formal negation $\neg G$ is also demonstrable.[17]

[17] Step 2 from Nagel & Newman, 2001, ch VII part B, p. 98 - 100

Suppose G is demonstrable in PM: that there does exist a proof of G in PM. Then, the formal negation of G,

$$\exists x \text{ Dem } (x,\ *\text{sub } (*n, *y, *n)) \qquad (\neg G)$$

stating *"There exists a proof of the statement G in PM"* must also be demonstrable. In other words, G is a demonstration of $\neg G$. Conversely, suppose that $\neg G$ is demonstrable, then there exists a proof that *"There exists a proof of G"*. And so G is demonstrable if, and only if, $\neg G$ is demonstrable.

We previously stated that if both a formula and its formal negation can be derived from the axioms of a formal system, then the formal system is not consistent. And conversely, if PM is a consistent formal system, neither G nor its negation is demonstrable, and so G is undecidable. As we have found a proof in PM of both G and $\neg G$, PM must be inconsistent. If we continue under the assumption that PM is consistent, it is then the case that G is an undecidable statement.

3.5.3 Step 3.

→ Show that although G is not formally demonstrable in PM consistent, it is nevertheless a true statement.[18]

We have determined that G is not formally demonstrable if PM is consistent. However, either G or $\neg G$ must be true. As we have seen, G states *"G is not demonstrable in PM"*. But we have in fact just proved G to be undecidable in PM, in particular, G has no proof inside PM. But, this is exactly what G asserts; and so it is clear to us that G asserts the truth. It is important to note that this truth value of G has been derived from outside the language of PM, and so does not constitute a demonstration of G within PM.

3.5.4 Step 4.

→ Show that since G is both true and formally undecidable (within PM), PM must be incomplete.[19]

We have already defined a formal system to be complete if every true statement that can be expressed in the system is formally deducible from the axioms by the rules of inference. If this is not the case, that is, if not every true statement expressible in the system is deducible, then the system is incomplete. It has just been established that the statement G, existing in PM, is true and

[18] Step 3 from Nagel & Newman, 2001, ch VII part B, pg 101 - 102
[19] Step 4 from Nagel & Newman, 2001, ch VII part B, p. 102 - 104

is not formally deducible within PM. It follows, therefore, that PM is an incomplete system, on the assumption that it is consistent.

Furthermore, PM is essentially incomplete, in that even if G were added as a further axiom, the augmented system PM + G would still not suffice to yield formally all arithmetical truths. Gödel shows that his method for producing undecidable formulas can be carried out no matter how often the initial system is enlarged.

3.6 Proving Gödel's Second Incompleteness Theorem

The second incompleteness theorem follows as an extension to the first incompleteness theorem, and relies upon constructions laid on in the above proof.

3.6.1 Step 5.

→ Construct a formula A of PM that represents the meta-mathematical statement *"PM is consistent"*.

The formula $A \supset G$ is formally undecidable inside PM. Furthermore, A is not demonstrable inside PM. Thus the consistency of PM cannot be established within PM.[20]

Having constructed the meta-mathematical statement, *"If PM is consistent, then it is incomplete"*, we wish to express this statement in the language of PM. The antecedent clause *"PM is consistent"* equivalently states *"There is at least one formula of PM that is not demonstrable inside PM"*. Via the process of Gödel numbering, this statement corresponds to the claim *"There is at least one formula whose Gödel number is $*y$ for which no proposed sequences of formulas whose Gödel number is $*x$ constitutes a proof of y inside PM"*, symbolically this produces the statement

$$(\exists y)(\sim \exists x)\ \text{Dem}\ (x, y). \tag{A}$$

The consequent clause *"it [PM] is incomplete"* is equivalent to stating of any true non-demonstrable formula X in PM, *"X is not a theorem of PM"*. But we have already constructed a formula for such a statement! The statement G says of itself, *"G is not a theorem of PM"*, and so our formal statement of G in PM, namely (G), can be used to represent the consequent clause.

We arrive at the full formula for the conditional statement,

$$(\exists y) \sim (\exists x)\ \text{Dem}\ (x, y) \supset \sim (\exists x)\ \text{Dem}\ (x,\ ^*\text{sub}\ (*n, *y, *n)). \tag{$A \supset G$}$$

[20] Step 5 from Nagel & Newman, 2001, ch VII part B p. 104 - 106

We now wish to show that A is not demonstrable in PM. To this end, suppose A is demonstrable in PM. Then, since the formula $A \supset G$ is demonstrable, by *modus ponens* the formula G is also demonstrable. But under the assumption that PM is consistent, G is formally undecidable and non demonstrable. Hence, the formula A is also non demonstrable in PM.

Hence the statement A is a formal expression inside PM of the meta-mathematical claim *"PM is consistent"*. Were we able to establish some chain of reasoning to reach the truth value of A, and if that sequence of steps could be mapped onto a sequence of formulas constituting a proof of A in PM, then A would be demonstrable in PM, and we would have deduced that PM is consistent. But, we have already shown that if PM is consistent, A is non demonstrable. And so we are forced to conclude that this sequence of steps to demonstrate A is impossible to find within PM, and that, finally, if PM is consistent, its consistency cannot be established by any metamathematical reasoning mirrored in the language of PM itself. Sadly for Hilbert, this conclusion immediately destructs his goal for a syntactical proof of the consistency of the numbers.

4 Computers in Mathematical Formalization

4.1 Introduction to Coq

Computer Proof Assistants are an interesting reaction to the mathematical crisis of the nineteenth century. They are evidence of mathematicians' persistence to find the singular correct answer to every problem. Proof Assistants come in various shapes and sizes; leading programs include Coq, Isabelle/HOL, and Mizar. The mathematical community has developed a list of 100 theorems to be formalized by a computer. So far the list is 93% complete.[21] Among the theorems are names you will have likely come across: the irrationality of $\sqrt{2}$, the Pythagorean theorem, Gödel's incompleteness theorem, the impossibility of trisecting an arbitrary angle, the fundamental theorem of calculus, that π is transcendental, that e is transcendental, the birthday problem, the law of cosines; to list a few.

For the purpose of this discussion, we will focus on Coq proof assistant. The first proof to be formalized in Coq was the Four Color Theorem in 2005 by Georges Gonthier, (the theorem was first conjectured by Francis Guthrie in 1852, and initially proved in 1976 by Appel and Haken). The theorem states that any map may be colored in less than 4 colors, such that no adjacent sections are of the same color. This was the second proof ever to be formalized by any computer proof system,

[21] Wiedijk, 2017

primarily since the proof is done by iterating over a large number of possible colorings and maps, and testing the truth of each case – a task one might consider computers were made specifically to perform.

So how does Coq actually formalize a proof? The user will begin by defining all necessary variables, functions and their respective types. Axioms and basic theorems of common arithmetic systems (the real numbers, the natural numbers, the integers etc.) are available for import from Coq's library. Then, the user will state and name what is to be proved with the relevant label (Theorem, Definition, Lemma, etc.). Coq will respond by listing the assumptions of the theorem, and stating the final goal. The user will then apply some 'tactic' to either the assumptions or the goal to manipulate them in some way. Coq has multiple libraries of common tactics, but users can also define a new tactic or sequence of tactics. Coq will execute then attempt to execute the user's command, and return the new goal or modified assumption. Eventually, the user will see that the current goal is shown to be true either inherently or based off of some listed assumption, and only then can the user use the tactic Qed which will command Coq to complete the proof. Coq will not allow you to formally end a proof it does not agree is fully formalized.

To see an example fragment of a Coq library, here is a selection of Coq's axioms for the Peano Natural numbers[22], and their corresponding statement in the common language of Peano[23].

Peano Axioms

1. $\forall x[Sx \neq 0]$
2. $\forall x, y[Sx = Sy \rightarrow x = y]$
3. $\forall x[x + 0 = x]$
4. $\forall x, y[x + Sy = S(x + y)]$
5. $\forall x[x \cdot 0 = 0]$
6. $\forall x, y[x \cdot Sy = (x \cdot y) + x]$
7. $\forall x[x^0 = S(0)]$
8. $\forall x, y[x^{Sy} = x^y \cdot x]$
9. For any axiom ψ above, and some function ϕ, $\phi(0) \wedge [\forall x(\phi(x) \rightarrow \phi(S(x))] \rightarrow \forall x \phi(x)$

Coq.Arith.PeanoNat

1. pred_0 : pred 0 = 0.
2. compare_succ n m : (Sn ?= Sm) = (n ?= m).
3. add_0_l n : 0 + n = n

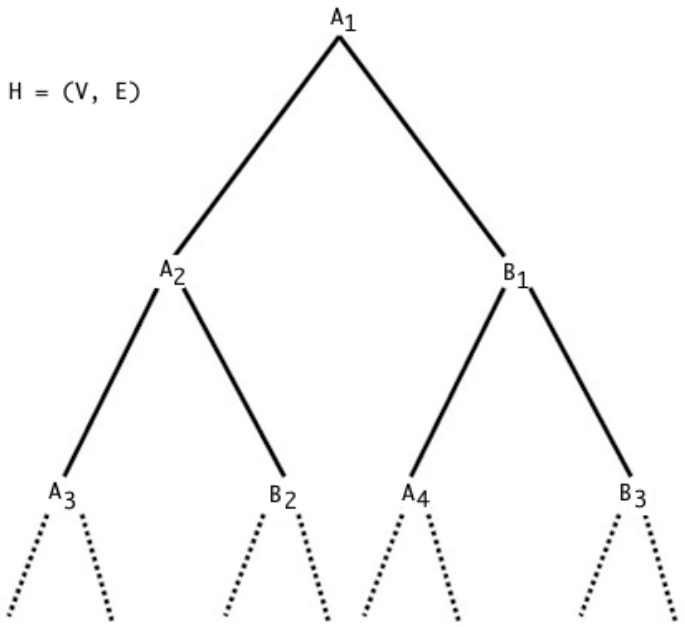
[22]Library Coq.Arith.PeanoNat.
[23]Goldstern & Judah, 1998

4. add_succ_l n m : (S n) + m = S (n + m)
5. mul_0_l n : 0 * n = 0
6. mul_succ_l n m : S n * m = n * m + m
7. pow_0_r a : a∧0 = 1.
8. pow_succ_r a b : 0<=b -> a ∧ (S b) = a * a∧b.
9. Theorem bi_induction :
   forall A : nat -> Prop, Proper (eq==>iff) A ->
   A 0 -> (forall n : nat, A n <-> A (S n)) -> forall n : nat, A n.

4.2 Russell's Paradox: presented in Coq

4.2.1 Thierry Coquand's Paradox of Trees

The best way to understand a Coq procedure and proof formalization is through example, we will present Russell's Paradox in Coq, in the form of Thierry Coquand's paradox of trees. First, we need to build an understanding of trees and how they relate to our discussion of self-referential sets. A tree is a type of graph, baring vertices connected by edges. Specifically, a tree is a non-circular graph where there is a unique path from the *root* vertex of the tree to any other vertex in the tree. We define a *subtree* to be any set of connected vertices. We permit trees to be infinite or recursive, allowing for a tree to have itself as a subtree. This gives us a meaningful form of visualizing sets that contain themselves.



In the tree $H = (V, E)$ above, the top-most vertex A_1 we will call the root of H. As the tree continues infinitely, each vertex branches to two new vertices, one A vertex and one B vertex. Each new A vertex can be considered a root of a subtree of H. Infinite trees, such as this one, have the special property that any subtree has a one-to-one correspondence mapping edges to edges and vertices to vertices with every other subtree, and with the whole tree. Imagine picking up one subtree by its root and placing it over the second subtree, starting with matching the two roots and working downwards over each vertex. As both trees continue infinitely and similarly, every vertex in the first subtree has an image vertex in the second subtree. Consequently, all subtrees are of equal order to each other, and to the whole tree.

> **Coquand's Paradox of Trees:** If we have a collection of trees, we are able to group them together under a common root to form a new tree. Every tree grouped under the new tree we will call an immediate subtree. Immediate subtrees are alike to proper subsets, in that an immediate subtree cannot contain every vertex in its encompassing tree. Furthermore, a tree will be called 'normal' or 'good' if it is not equal to any of its immediate subtrees, that is, if it does not experience this infinite recursion. Now, suppose the new tree R that contains all normal trees, and only normal trees. Is R itself normal?
>
> Suppose R is normal. By the construction of R having all normal trees as subtrees, R is a subtree of R. But, the definition of normal implies that R is not equal to any of its immediate subtrees. So R is not normal. If R is not normal, since R contains only normal subtrees, R is not a subtree of R. But, by definition of normal, if R is not a subtree of R, R is normal. However, we already deduced that R could not be normal. This leads to a logical contradiction.[24]

4.2.2 Proof in Coq Script

Section Russell.[25]

Set Implicit Arguments.

Variable set : Set.

Variable name : Set -> set.

[24] Coquand, 1992

[25] Exact proof in Coq by Altenkirch, 2009

```
Variable El : set -> Set.

Axiom reflect : forall A:Set, A = El (name A).

Inductive Tree : Set :=
    span : forall a : set, (El a -> Tree) -> Tree.

Definition elem (t : Tree) (u : Tree) : Prop
    := match u with
    | span A us => exists a : El A , t = us a      end.

Definition Bad (t : Tree) : Prop
    := elem t t.

Definition GoodTree : Set
    :=  t : Tree |   Bad t .

Definition goodTree : set
    := name GoodTree.

Definition getTreeAux : forall A:Set,(GoodTree = A) -> A -> Tree.
    unfold GoodTree.
    intros A eq.
    rewrite <- eq.
    intros gt.
    destruct gt.
    exact x.
    Defined.

Definition getTreePropAux : forall (A:Set) (p: GoodTree = A)(a : A), ~ Bad (getTreeAux p a).
    unfold GoodTree.
    intros A eq.
    dependent inversion eq.
    intros.
    simpl.
    destruct a.
    exact n.
    Defined.

Definition getTree : El goodTree -> Tree.
    unfold goodTree.
    apply getTreeAux.
    apply reflect.
    Defined.

Lemma getTreeProp : forall g : El goodTree,   Bad (getTree g).
    apply getTreePropAux.
    Qed.

Definition mkGoodAux : forall A:Set, GoodTree = A -> forall t:Tree, ~ Bad t -> A.
    intros A eq.
    rewrite <- eq.
```

```
    intros t nb.
    exists t.
    exact nb.
    Defined.

Lemma mkGoodPropAux : forall (A:Set)(eq:GoodTree = A), forall (t : Tree)(p: ∼ Bad t), t =
getTreeAux eq (mkGoodAux eq t p).
    intros A eq.
    dependent inversion eq.
    intros t p.
    simpl.
    Reflexivity.
    Qed.

Definition mkGood (t : Tree)(p: ∼ Bad t) : El goodTree.
    apply mkGoodAux.
    apply reflect.
    Defined.

Lemma mkGoodProp : forall (t : Tree)(p: ∼ Bad t), t = getTree (mkGood t p).
    apply mkGoodPropAux.
    Qed.

Definition russell : Tree
    := span getTree.

Lemma bad_Imp_Good : Bad russell -> ∼ (Bad russell).
    unfold Bad at 1.
    unfold russell at 2.
    unfold elem at 1.
    intro H.
    elim H.
    intros.
    rewrite H0.
    apply getTreeProp.
    Qed.

Lemma good_Imp_Bad: ∼(Bad russell) -> Bad russell.
    intros.
    unfold Bad.
    unfold russell at 2.
    unfold elem.
    unfold GoodTree.
    exists (mkGood russell H).
    apply mkGoodProp.
    Qed.

Lemma goodRussell : ∼(Bad russell).
    intro H.
    apply bad_Imp_Good.
    exact H.
    exact H.
```

Qed.

Lemma badRussell : Bad russell.
apply good_imp_Bad.
apply goodRussell.
Qed.

Lemma paradox : False.
apply goodRussell.
apply badRussell.
Qed.

4.2.3 Running a Coq Proof

To display every exchange between program and user in crafting a proof formalization in Coq message would be excessive. To give at least some sense of the user interacts with Coq, consider the Lemma bad_Imp_Good shown below. On the left is the proof script, lines that have been executed are left blue. On the right are the goals and assumptions Coq outlines after the execution of the most recent step. Assumptions are listed above the dashed line, and the formula(s) to be proved is shown the line below. The tactics we will see in the proof below are some of the most commonly used tactics in Coq proofs.

```
Lemma bad_Imp_Good : Bad russell -> ~ (Bad russell).
  unfold Bad at 1.
  unfold russell at 2.
  unfold elem at 1.
  intro H.
  elim H.
  intros.
  rewrite H0.
  apply getTreeProp.
Qed.
```

```
1 subgoal, subgoal 1 (ID 9)

  set : Set
  name : Set -> set
  El : set -> Set
  ============================
  Bad russell -> ~ Bad russell
```

The underlying types of variables are listed as assumptions, and our goal is stated.

```
Lemma bad_Imp_Good : Bad russell -> ~ (Bad russell).
  unfold Bad at 1.
  unfold russell at 2.
  unfold elem at 1.
  intro H.
  elim H.
  intros.
  rewrite H0.
  apply getTreeProp.
Qed.
```

```
1 subgoal, subgoal 1 (ID 10)

  set : Set
  name : Set -> set
  El : set -> Set
  ============================
  elem russell russell -> ~ Bad russell
```

The *unfold* tactic used in the next three steps uses the definition of the specified variable to rewrite that variable in the current goal. If the variable occurs more than once, the keyword *at* is used to specify which occurence we wish to alter.

```
Lemma bad_Imp_Good : Bad russell -> ~ (Bad russell).
  unfold Bad at 1.
  unfold russell at 2.
  unfold elem at 1.
  intro H.
  elim H.
  intros.
  rewrite H0.
  apply getTreeProp.
Qed.
```

```
1 subgoal, subgoal 1 (ID 11)

  set : Set
  name : Set -> set
  El : set -> Set
  ============================
  elem russell (span GoodTree getTree) -> ~ Bad russell
```

```
Lemma bad_Imp_Good : Bad russell -> ~ (Bad russell).
  unfold Bad at 1.
  unfold russell at 2.
  unfold elem at 1.
  intro H.
  elim H.
  intros.
  rewrite H0.
  apply getTreeProp.
Qed.
```

```
1 subgoal, subgoal 1 (ID 12)

  set : Set
  name : Set -> set
  El : set -> Set
  ============================
  (exists a : El GoodTree, russell = getTree a) ->
  ~ Bad russell
```

```
Lemma bad_Imp_Good : Bad russell -> ~ (Bad russell).
  unfold Bad at 1.
  unfold russell at 2.
  unfold elem at 1.
  intro H.
  elim H.
  intros.
  rewrite H0.
  apply getTreeProp.
Qed.
```

```
1 subgoal, subgoal 1 (ID 13)

  set : Set
  name : Set -> set
  El : set -> Set
  H : exists a : El GoodTree, russell = getTree a
  ============================
  ~ Bad russell
```

As our goal contains an implication, *intro H* brings the antecedent to our list of assumptions under the label H.

```
Lemma bad_Imp_Good : Bad russell -> ~ (Bad russell).
  unfold Bad at 1.
  unfold russell at 2.
  unfold elem at 1.
  intro H.
  elim H.
  intros.
  rewrite H0.
  apply getTreeProp.
Qed.
```

```
1 subgoal, subgoal 1 (ID 14)

  set : Set
  name : Set -> set
  El : set -> Set
  H : exists a : El GoodTree, russell = getTree a
  ============================
  forall x : El GoodTree, russell = getTree x -> ~ Bad russell
```

The tactic *elim H* introduces the universal quantification of H as an implication in our goal. This seems like a step backwards, but will bring our goal closer to an already prove lemma.

```
Lemma bad_Imp_Good : Bad russell -> ~ (Bad russell).
  unfold Bad at 1.
  unfold russell at 2.
  unfold elem at 1.
  intro H.
  elim H.
  intros.
  rewrite H0.
  apply getTreeProp.
Qed.
```

```
1 subgoal, subgoal 1 (ID 16)

  set : Set
  name : Set -> set
  El : set -> Set
  H : exists a : El GoodTree, russell = getTree a
  x : El GoodTree
  H0 : russell = getTree x
  ============================
  ~ Bad russell
```

Again, *intros* will bring the antecedent of the implication in our goal to the list of assumptions under the arbitrary name $H0$. It is different to the *intro* tactic in that it wall attempt to draw all possible assumptions, rather than a single assumption H.

```
Lemma bad_Imp_Good : Bad russell -> ~ (Bad russell).
  unfold Bad at 1.
  unfold russell at 2.
  unfold elem at 1.
  intro H.
  elim H.
  intros.
  rewrite H0.
  apply getTreeProp.
Qed.
```

```
1 subgoal, subgoal 1 (ID 17)

  set : Set
  name : Set -> set
  El : set -> Set
  H : exists a : El GoodTree, russell = getTree a
  x : El GoodTree
  H0 : russell = getTree x
  ============================
  ~ Bad (getTree x)
```

The *rewrite H0* command is similar to the *unfold* tactic, in that our goal is rewritten using what is given by $H0$.

```
Lemma bad_Imp_Good : Bad russell -> ~ (Bad russell).
  unfold Bad at 1.
  unfold russell at 2.
  unfold elem at 1.
  intro H.
  elim H.
  intros.
  rewrite H0.
  apply getTreeProp.
Qed.
```

```
No more subgoals.
(dependent evars: (printing disabled) )
```

Note that the lemma *getTreeProp* is has already been declared above. The tactic *apply getTreeProp* asks Coq to determine if this lemma can decide the truth of or simplify our current goal. Coq responds that this lemma does in fact prove the current and final goal, which completes the proof.

```
Lemma bad_Imp_Good : Bad russell -> ~ (Bad russell).
  unfold Bad at 1.
  unfold russell at 2.
  unfold elem at 1.
  intro H.
  elim H.
  intros.
  rewrite H0.
  apply getTreeProp.
Qed.
```

```
bad_Imp_Good is defined
```

With the command *Qed*, we formally define the lemma in the proof environment.

4.2.4 Unpacking the Proof of Russell's Paradox

Coq attempts to prevent self referential sets, and thus a proof of Thierry's paradox requires defining the following: the variable set that is of type Set; the function name mapping Set onto set; and the function El mapping set onto Set. This will allow us to define the base unit of Thiery Coquand's paradox: the Tree. We define the Tree as a Set, all of whose elements are also Trees. Furthermore, we define a *Bad* Tree: a Tree is Bad if it has itself as an element. This provides the basis to define a *GoodTree*: a Set of Trees, none of which are the GoodTree itself. Finally, we define the set *goodTree*: a set whose elements are all GoodTrees.

The proof then has a sequences of lemmas and definitions leading up to the definition of *getTree* and the lemma *getTreeProp*. As is often the case in set theoretic proofs, we wish to target a specific element in the set; to this end, getTree defines a particular Tree in a goodTree. getTreeProp says that no Tree in a goodTree is an element of any other Tree in a goodTree; or equivalently, for any Tree g in a goodTree, no tree in a goodTree is a subtree of g. Then follows another sequence of lemmas and definitions to define the property mkGood and the lemma mkGoodProp. mkGood defines a particular Tree in a goodTree that is not Bad. Then, mkGoodProp says that all Trees that are not Bad are elements of a goodTree. These two definitions and two lemmas allow us to begin crafting the truly paradoxical tree *russell*: russell is a Tree whose elements are defined to be exactly each and every Tree that is an element of each and every goodTree.

Then we begin the paradoxical spiral of this proof: Lemma l1 states that if russell if Bad, then

russell is not Bad. We have defined that if russell is Bad, it contains itself as an element. But russell only contains Trees in a goodTree, and all Trees in a goodTree are GoodTrees, not Bad. So russell must not be Bad if russell is Bad. Lemma l2 states that if russell is not Bad, then russel is Bad. We know that if russell contains all Trees in a goodTree. But if russell is not Bad, then it does not contain itself as an element, and so itself is not an element of a goodTree, but all GoodTrees are an element of a goodTree. And so russell must be Bad if russell is not Bad. Either russell is Bad or it is not; the lemma goodRussell relies on l2 to infer that russell is not Bad, and the lemma badRussll relies on l1 to infer that russell is Bad. Finally, we wish to provide proof of a falsity, for this confirms the paradox of the Tree russell. We apply our derivation that russell is good, which implies by good_Imp_Bad that russell must be Bad. We apply our derivation that russell is Bad, which implies by bad_Imp_Good that russell must be not be Bad. This contradicts the assumption we just introduced that russell is Bad, and thus provides a proof of False.

4.3 Computer Formalization: A continuation of Hilbert's Plan

Computer proof assistants are a direct echo of Hilbert's *non ignorabimus*: his determinism to formalize mathematics, to be thorough in our work so as to be sure we do not make assumptions about mathematics based off of human intuition where there are no mathematical grounds to do so. But how could this ideology continue after Gödel's work and his destruction of Hilbert's plan? Surely, what we learned from Gödel is that in certain cases, we will come across true statements that cannot be proved using the axioms of the particular formal system. Furthermore, Gödel's theorems directly conflict Hilbert's goal to find an algorithm determining all proofs within the natural numbers by showing that no such algorithm exists. And yet computer proof assistants remain a valuable computational tool aiding in proof formalization.

So as to be accurate in their formalization, computer proof assistants mirror the boundaries of our formal systems. Computer proof assistants do the job they are designed for very well; they do not use human intuition, and it is for this very reason we trust them so thoroughly to check our proofs. And yet, this trait can also be seen as a downfall of proof assistants; they do nothing to overcome the incompleteness of our formal arithmetic systems. Humans are able to think from outside formal systems, to construct metamathematical statements more useful than their formulaic counterparts. As we have seen with the liar paradox and Gödel statements, there are cases in which we are able to recognize the truth of a proposition from outside the system without forming a proof of its truth

within the system.

Consequently, we will never be able to hand over full control to computer automated proof assistants. They do not have the intuition and ability to look meta-mathematically, taking a step backwards from the native language of the system; this task must be left to the human user. We may eventually formalize every proof mathematicians have found within each formal system, but a computer cannot take over the task of the human. And this, I believe, is something David Hilbert would be thoroughly glad to hear.

5 Conclusion

References

Aharoni, R. (2015). Mathematics, poetry, and beauty. New Jersey: World Scientific.

Altenkirch, T. (2009, November). L16 Russell's paradox [Coq Proof Assistant]. University of Nottingham. Retrieved from sympa.inria.fr/sympa/arc/coq-club/2008-11/msg00109.html

Avigad, J., & Reck, E. H. (2001). Clarifying the nature of the infinite: the development of metamathematics and proof theory. Carnegie Mellon. Retrieved from www.cmu.edu/dietrich/philosophy/docs/tech-reports/120_Avigad.pdf

Copeland, B. J., & Posy, C. J. (2013). Computability: Turing, Gödel, Church, and Beyond. MIT Press.

Coquand, T. (1992). The paradox of trees in Type Theory. INRIA, and University of Göteborg/Chalmers. Retrieved from www.cse.chalmers.se/ coquand/tree.ps

Farlow, J. (2008). Section 2.6: Cantor's Theorem. In A Taste of Pure Mathematics. University of Maine. Retrieved from www.math.umaine.edu/ farlow/sec26.pdf

Franzen, T. (2005). Gödel's Theorem: An Incomplete Guide to Its Use and Abuse. Wellesley, Massachusetts: A K Petters.

Goldstern, M., & Judah, H. (1998). The Incompleteness Phenomenon: A New Course in Mathematical Logic. Natick, Massachusetts: A K Petters.

Hamilton, N. T., & Landin, J. (1961). Set theory and the structure of arithmetic. Boston: Allyn and Bacon.

Library Coq.Arith.PeanoNat. (2017, October). Retrieved from coq.inria.fr/distrib/current/stdlib/Coq.Arith.PeanoNat

Nagel, E., & Newman, J. R. (2001). Gödel's Proof (Revised Edition). New York University Press.

Mastin, L. (2010). 19th Century Mathematic. Retrieved from www.storyofmathematics.com/19th.html

Part VI Mathematicians. (2008). In Princeton Companion to Mathematics (p. 94). Princeton University Press.

Paulin-Mohring, C. (2013). Introduction to the Coq proof-assistant for practical software verification. Retrieved from www.lri.fr/ paulin/LASER/course-notes.pdf

Smith, P. (2013). An Introduction to Gödel's Theorem. Cambridge University Press.

Wiedijk, F. (2017, October). Formalizing 100 Theorems. Retrieved from www.cs.ru.nl/ freek/100/

Sarah Dennis
Mathematical Modeling II
Philip Ording
Spring Conference Paper
May 12 2017

An Approach to the Small World Problem:

Analysing the spread of diseases and computer viruses through graph theory

1 Introduction

Increasing globalisation and the high level of connectedness we experience on a day to day basis, puts humanity in a complex position. The problem of modeling the human network is at the forefront of understanding the mechanics of spread (of disease, fashion trends, computer viruses, rumors, etc.). Our global interpersonal network poses a great challenge to model and understand - hence the term small world problem. To unravel the problem of modeling spread in a small world, we can apply the basic principles of graph theory to enhance our understanding of pre-existing mathematical models for spread.

Graph theory is used to analyse networks and connections. These graphs are different from charts; graphs feature points (also referred to as nodes or vertices) that can be connected by any number of edges. There is no scale or specific ordering of vertices, so long as the integrity of the connections is preserved. We can, theoretically, make a graph that models the connections (physical or virtual) of the human population. This would be a graph of incredibly large scale, changing every second, and likely not very precise. However, we can analyse the properties and behaviors of graphs of a similar type, but of smaller scale, and make predictions of patterns occurring in our larger network.

Small world graphs take an interesting form, with their characteristics rooted partially in random graphs and partially in lattice graphs. Their is some equilibrium between these two extremes where our particular small world graph lies. Understanding the characteristics of our network through graph theory allows

us to build on preexisting models for network based behavioral patterns. We can draw connections between the features of our network that are detrimental and those that are beneficial to our population's well being. For our purposes, suppose that we would like to analyse how disease and computer virus spread is dependent on the structure of our network, so that we may understand ways in which our network could be altered and improved.

In terms of disease spread, we can look at the SIR model for a particular disease that has already had an epidemic outbreak and create a graph that approximates the network through which the disease spread. This will allow us to suppose the outcome of that same disease spreading in a differently structured network, or a different disease spreading in the same network, hence we are able to efficiently and effectively allocate resources for combating disease. This process will also enable us to identify the elements of each network structure that facilitated or held back the spread of the disease, and see the ways in which a certain disease behaves depending on the network structure. In terms of computer viruses, analysing how a virus has spread allows us to understand the robustness of our online network, highlighting the virtual connections that are most easily compromised by hacker technology, so that we may take precautions for safer computer use in the future. In total, the process of analysing the characteristics of a small world graph will allow us to recognise the impacts of connectedness on the spread of diseases and the problems with uniformity for the spread of computer viruses.

2 An Introduction to Graph Theory

For us to discuss and compare different graphs in a meaningful way, we require a basic level of graph theory terminology. Here we will introduce the definitions of graph theory concepts and explain their relevance to our discussion of disease and computer virus spread:

A *graph G* consists of a nonempty set of elements, called *vertices*, and a list of unordered pairs of these elements, called *edges*. The set of vertices of the graph *G* is called the *vertex set* of *G*, denoted by *V(G)*, and the list of edges is called the *edge list* of *G*, denoted by *E(G)*. If *v* and *w* are vertices of *G*, then an edge connecting the two is said to be *incident* on *v* and on *w*. If an edge is incident on *v* and *w*, the two vertices are said to be *adjacent* to one another. The number of vertices in *V(G)* is termed the *order* of the graph (*n*), and the number of edges in *E(G)* is termed its *size* (*M*). The *degree* of a graph, denoted *k*, is the average number of edges incident on *v* for every vertex in *V(G)*.

For our application to the spread of disease, we will consider each person as a vertex. Defining a connection or an edge is at the heart of the complexity of modeling our network, but for our particular application to disease spread, we can count physical contact as a connection, and assume that nothing but time prevents disease from transferring through every available edge. For computer virus spread, we can either consider a person to be their email server, or their physical PC computer, depending on how a specific computer virus acts. Edges will most commonly be defined by connection through email contacts or through overlaps in website activity (such as online video providers, shopping sites, etc.).

The *Shortest Path Length* of a graph is defined as the minimum number of edges that must be traversed in order to reach vertex *j* from vertex *i*. The *Characteristic Path Length* (L) of a graph is the median of the means of the shortest path lengths connecting each vertex $v \in V(G)$ to all other vertices. That is, calculate $d(v, j)\ \forall j \in V(G)$ and find $đ_v$ for each *v*. Then define *L* as the median of $\{đ_v\}$.

The *subgraph* of a graph *G*, denoted *G'*, is a graph whose vertices *V(G')* and edges *E(G')* are subsets of *V(G)* and *E(G)* respectively. The *neighbourhood* $\Gamma(v)$ of a vertex *v* is the subgraph that consists of the vertices adjacent to *v* (but not including *v* itself). Two vertices are *neighbours* if they exist in the same neighbourhood.

The *clustering coefficient* γ_v of a neighborhood Γv characterises the extent to which vertices adjacent to any vertex *v* are adjacent to each other:

$$\gamma_v = |E(\Gamma_v)| / \begin{pmatrix} k_v \\ 2 \end{pmatrix}.$$

The *clustering coefficient of a graph G*, denoted γ, is $\gamma = \gamma_v$ averaged over all $v \in V(G)$. Hence $\gamma = 1$ would imply that the corresponding graph consisted of n/(k + 1) disconnected, but individually complete, subgraphs, and $\gamma = 0$ would imply that no neighbour of *any* vertex *v* is adjacent with any other neighbour of *v*.

Path length and clustering coefficient are important concepts for our comparison of lattices and random graphs, and hence for our approximation of the properties of a small world graph. An understanding of neighborhoods is essential to epidemic prevention. Consider two neighborhoods in your personal network: groups of people you know who all know one another. If you, or any person in either neighborhood, contracts a disease, and we assume physical contact between members of each neighborhood, every person in both neighborhoods will become infected. This concept will be further developed with an understanding of reproduction rate.

The *range* of an edge, denoted *R(i, j),* is the length of the shortest path between *i* and *j* in the absence of that edge. An edge *(i, j)* with a range *R(i, j) = r* is called an *r-edge*. An r-edge with r>2 is called a *shortcut*, meaning it connects two vertices that would otherwise be spanned by more than 2 edges.

Shortcuts are essential to controlling disease spread. Consider the same two neighborhoods in your network - you represent the shortcut between these two neighborhoods, and will be the way in which a disease is able to travel from one to the other.

3 Random Graphs, Lattice Graphs & Small World Graphs

3.1.1 Defining Random Graphs and Lattice Graphs

A *random graph* of order n is a vertex set consisting of n vertices, and an edge set that is generated in some random fashion. Random graphs generally have clustering coefficient $\gamma = k/n$. There is currently no closed-form approximation for random graph characteristic path length, but we can assume a value close to $L \sim ln(n)/ ln(k)$.

A *lattice graph* is a simple graph where any vertex v is joined to its lattice neighbours so that each vertex v has the same order k. A 1-lattice with even $k \geq 2$ has characteristic path length $L = \frac{n(n+k-2)}{2k(n-1)}$ and clustering coefficient $\gamma = \frac{3(k-2)}{4(k-1)}$. While lattice graphs can vary widely, each is highly structured and ordered: giving every vertex the same order, and having a standard pattern for connections to be distributed on. An essential characteristic of a lattice graph is that they have no shortcuts.

3.1.2 Comparing Random and Lattice Graphs

Consider the network we live in today, clearly the lattice graph cannot accurately model our connections. Technological innovations in transport (affecting disease spread) and communications (affecting computer virus spread) have made shortcuts such an essential part of our day to day lives. And yet our network of connections is not fully random either, for we can certainly identify certain social circles where clustering is high. As such, we assume that our network structure resides at some medium between the two extremes of structured and random. Where this medium is, we cannot be precisely sure of, and it is always changing. However, knowing that our network has characteristics being drawn from these two basic and well studied graph structures will give us insight into what we can expect from a model of our network.

3.2 Beta Model of Rewiring

The next step in understanding what lies between a lattice and random graph is to analyse the β-model of rewiring. The algorithm for the β-model starts with a perfect 1-lattice, and then randomly *rewires* the edges of the lattice with some probability β. Each edge in the graph is to be considered for rewiring exactly once: for some edge i connected to its nearest neighbor $i + 1$, take some random deviate r; if $r < \beta$, $(i, i + 1)$ is rewired to another vertex j (chosen randomly from the vertex set), otherwise, the edge is unchanged. At the end of this process, no new edges have been created. Hence, when $\beta = 0$, the graph remains a 1-lattice, and when $\beta = 1$, the result is a fully random graph of the same order. This is an essential process to consider, since there must be some β that will give a graph equivalent to the network we live in.

We noted that a key characteristic of a lattice graph is that they have no shortcuts due to the fact that each vertex has exactly the same order; each vertex in lattice graph will also have have the same shortest path length (and it will be high). In a random graph, shortcuts are frequent since there is no regulation on how many vertices a single edge can connect. As such, lattices have higher clustering coefficients and longer mean path lengths, and random graphs have lower clustering and shorter mean path length. Thus our prediction for the beta rewiring model is that path length and clustering coefficient will both decrease as we approach a random graph.

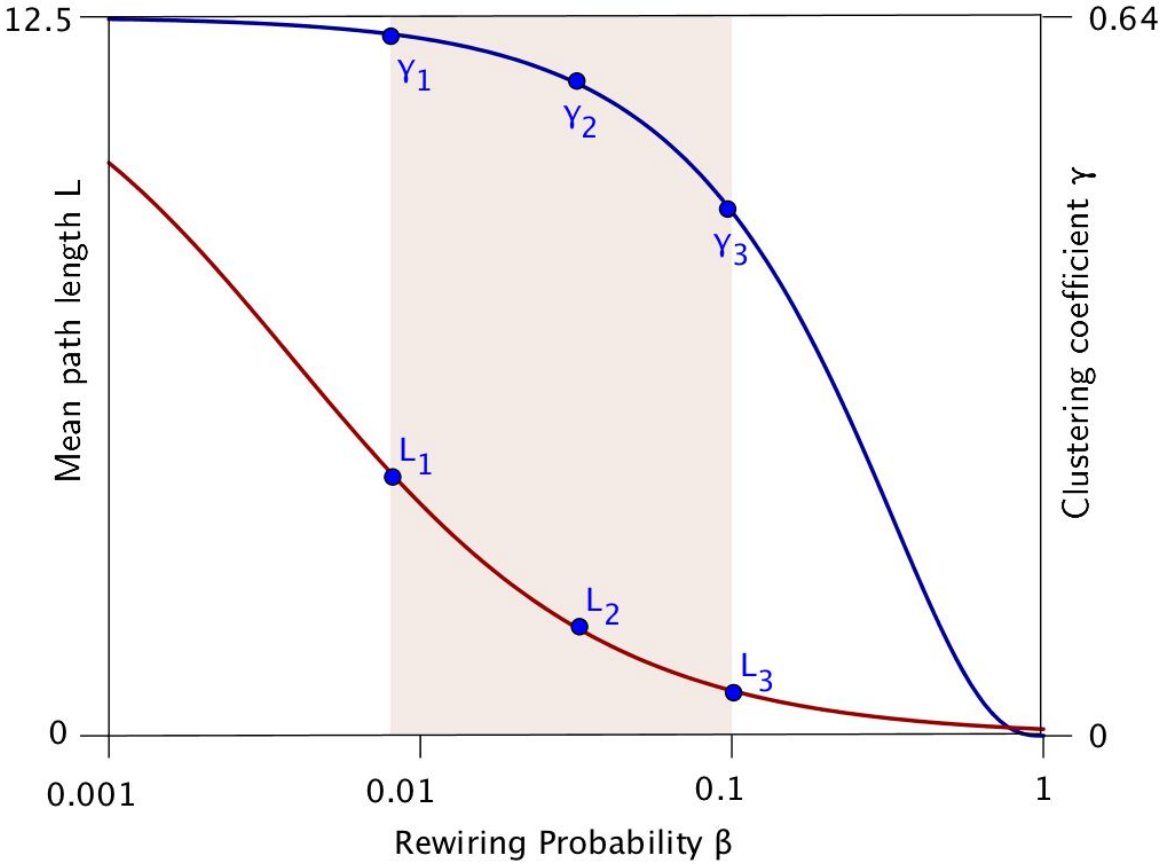


Fig 1. Graphing Mean Path Length (red) and Clustering Coefficient (blue) through the beta rewiring process.

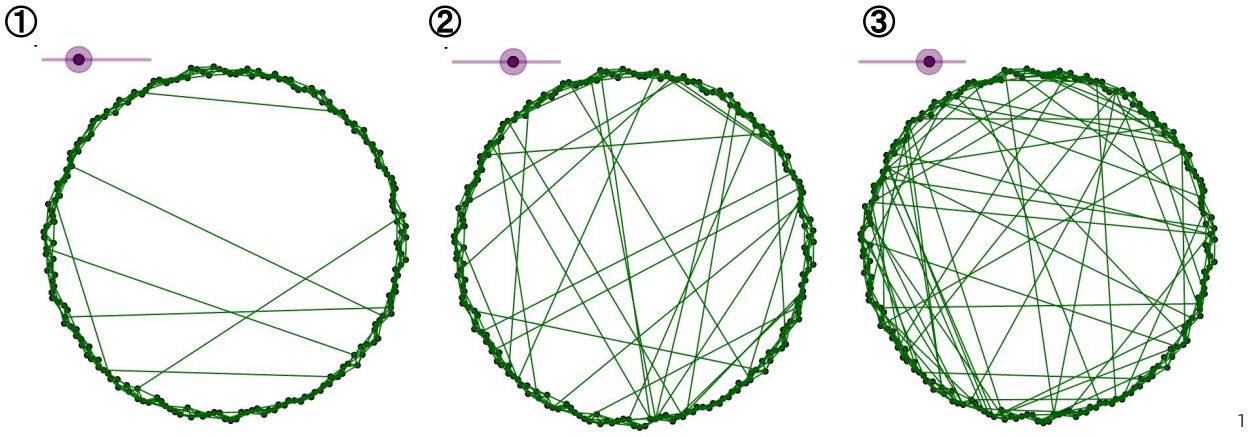


Fig 2. A graph undergoing the process of beta rewiring.

[1] Graphs captured and chart modified (via Geogebra) from mathinsight.org/small_world_network

	① Maximum	② Median	③ Minimum
Clustering Coefficient	$\gamma_1 = 0.63$	$\gamma_2 = 0.58$	$\gamma_1 = 0.47$
Mean Path Length	$L_1 = 4.55$	$L_2 = 1.88$	$L_3 = 0.79$

Fig 3. Table showing the values for max, median, and min of mean path length and clustering coefficient

Fig 1. shows how mean path length and clustering coefficient change with the beta rewiring process. The shaded region represents the values of β that will result in a small world graph. The points on each curve correspond to the position in the rewiring process of each graph in *Fig 2.* The graphs in *Fig 2,* are possible outcomes for graphs with $n = 200$ in transition from a lattice of $k_v = 8$ to a random graph, (note that the value of p shown in the top left corner of each graph is equivalent to β). The 3 stages correspond respectively to the maximum, median and minimum values of mean path length and clustering coefficient for a small world graph. The specific values for these parameters (averaged over many trials for β) are shown in *Fig 3.*

3.3 Defining a small world graph

We can now properly define what it means for our network to be a *small world graph*. The beta rewiring model (*Fig 1.*) shows that a small world graphs has the following properties…

1. a characteristic path length comparable to the shortest path length for a random graph of that size: $L \approx L_{random}(n,k)$.
2. a clustering coefficient greater than we would expect for an equivalent sized random graph: $\gamma \gg \gamma_{random} \approx k/n$

The small world graph then has an order k somewhere between that of a random graph (w.r.t path length) and that of a 1 lattice (w.r.t clustering), and some value β

being the probability of edge reassignment as transition occurs from a lattice graph to a random graph such that $0.01 \leq \beta \leq 0.1$.

4 Modeling Spread in a Small World

Transitioning into analysing disease spread, consider the expected growth curve of any disease. We expect to be some variation on an s-shaped logistic curve:

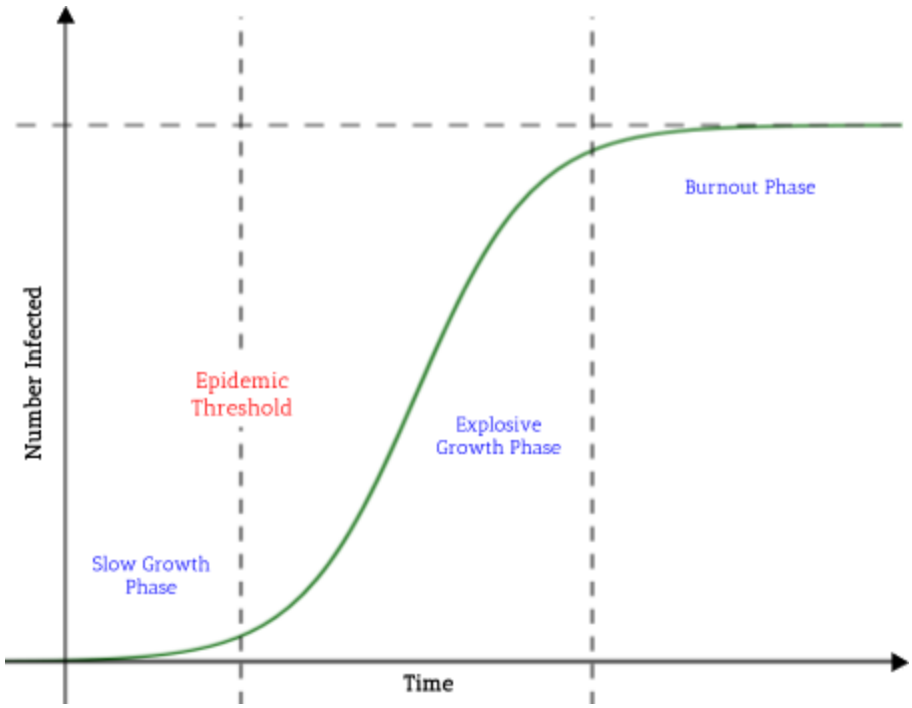


Fig 4. Chart for expected growth curve of disease

In this curve we can pick out three distinct phases: Slow growth, Explosive growth, and Burnout. The transition from slow growth to explosive growth occurs when the a disease's reproduction rate (the number of people a person infects) exceeds one, as the disease has began spreading at an increasing rate; this is the mathematical definition of an epidemic. The approach of the reproduction rate to 1 is called the *threshold* of an epidemic. In order to prevent an epidemic, the reproduction rate must be kept below its threshold.

To calculate the reproduction rate for a disease, we can use the SIR model. The effect of the reproduction rate and its importance for controlling spread can be analysed using our knowledge of a graph's neighborhoods.

4.1 The SIR model for disease spread

The SIR model was developed over 80 years ago by William Kermack and A.G. McKendrick, and is still at the foundations of disease modeling today. The model is most utilitarian when used to perfect our understanding of past epidemics. It takes variables that are difficult to accurately approximate during the course of a disease, and as such, we make use of the SIR model to prepare for combating a similar disease epidemic in the future.

The SIR model is based on the idea that we can break up a network into three categories of people by their infection status, making up the acronym of the model:

S → Susceptible: an individual who is not infected but is vulnerable to infection

I → Infectious: an individual who is infected and is capable of infecting others

R → Removed: an individual who is recovered/dead and poses no further threat

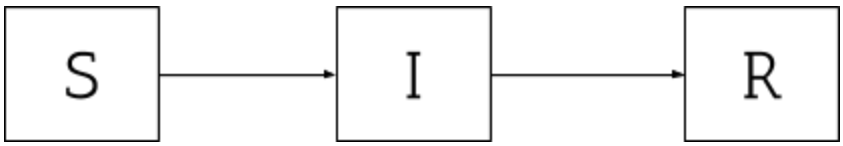


Fig 5. Basic compartmental diagram for the SIR model

We can then generate differential equations to model the change in population of each category as the disease spreads.

$$\frac{dS}{dt} = -\frac{\beta SI}{N}$$
$$\frac{dI}{dt} = \frac{\beta SI}{N} - \mu I$$

$$\frac{dR}{dt} = \mu I$$

$S(t)$, $R(t)$ and $I(t)$ represent the population of each category at time ***t***. The sum of each category gives the total population N at time ***t***:

$$N = S(t) + I(t) + R(t).$$

The average probability of infection β', is given by

$$\beta' = pc$$

where ***p*** is the probability of infection after exposure to an infected person, and ***c*** is the per capita contact rate. Then,

$$\frac{\beta SI}{N}$$

represents the number of people infected at time ***t***. This value causes the decrease in the susceptible population and is attributed to the increase in the infected population. The infected population also has the decrease in population from the death rate μI, where μ is the per capita death rate. Note that these equations do not account for those that recover and become susceptible once again – we assume a level of immunity that prevents reinfection.

As previously mentioned, these differential equations require variables that are difficult to calculate for the a disease that is in the process of spreading; specifically, we need values for β', **p, c,** and μ that will not be obvious until the disease has run its full course. Hence, these values are best found by fitting each differential equation to real data that has been gathered in the past, and extracting the variables in this fashion.

As shown by our examination of the S-shaped logistic curve, one way to understand the behavior of an epidemic is through reproduction rate $\mathcal{R}$. Once values for β', and μ have been determined, we are able to calculate $\mathcal{R}$ as the ratio between the average probability of infection and the death rate:

$$\mathcal{R} = \beta'/\mu$$

Obtaining a value for reproduction rate, the SIR model allows us to quantify the effectiveness of the disease in terms of spreading capability so that various diseases can be easily compared. We will delve deeper into how reproduction rate can aid our understanding and improve predictions of disease spread through analysing the spread of disease through random and lattice graphs.

4.2 The SIR Model Applied to Graph Theory

Given that a small world graph is halfway between a random graph and a lattice graph, we can approximate the behavior of diseases in each extreme scenario with respect to the SIR model. Our network has the characteristic path length of a random graph, and the clustering coefficient of a lattice graph; we can consider the ways in which these characteristics contribute to the behavior of spread in lattice and random graphs, and create the middle ground scenario that would represent the behavior of disease and computer virus spread in our small world network.

4.2.1 SIR Model in Random Graphs

First, we wish to relate the behavior of a disease under the SIR model to what we understand about connectivity in a random graph. To this end, consider each person a vertex on a random graph – we are ignoring the structure of our social network by connecting people in an entirely random fashion. And, have each connection represent interactions that are appropriate to the transfer of the particular disease; in most cases this will be close physical interaction.

It then follows that the per capita connection rate c is in fact identical to a random graph's average number of neighbors. Then, as each of the vertices in the neighborhood of the infectious can become infected too, with some probability p, the reproduction rate is also highly correlated to the average number of neighbors.

Now, we can make the connection that the epidemic threshold $\mathcal{R} = 1$ is strongly related to the connection of two neighbourhoods in a random network.

When a shortcut is made between two vertices and two neighborhoods are suddenly within reach where before they were vastly separated, the number of people one is connected to has been dramatically increased. In upping the contact rate c, the dependence on infection probability p for disease transfer is decreased – more damage can be done with a smaller infection probability: the probability of infection will be calculated on a single vertex as many times as its number of infected neighbors, and even a low infection rate will eventually be unable to keep the reproduction rate below one and prevent the disease from reaching epidemic.

4.2.2 SIR in Lattice Graphs

An alternate comparison can be made between the SIR model and a lattice graph. In a lattice model of our population, the high clustering coefficient implies that a spreading disease is continually being forced back into the already infected population – the connections of those who become infected are likely already infected and cannot be infected again, and so as the number of infectives increases, the respective damage they are able to do in terms of disease spread is less and less. Consider a 1 lattice with a growing neighborhood of infectives. This neighborhood consists of two types of points, those in the interior of the cluster who cannot infect any susceptibles and those on the exterior of the cluster who can infect and who form the disease front. Since growth is only considered in one dimension, the size of the disease front remains fixed even as the infected neighborhood grows. Thus, the reproduction rate for the infected population is decreasing as the infection spreads.

This conclusion implies that for a disease confined to spread in only a limited number of dimensions (lattice-style), only the most infectious diseases will develop into true epidemics. And even then, spread will be slow and creeping rather than explosive, giving time for adjustment and precautionary action. So, the same disease spreading in a lattice will tend to infect far fewer people than in a random graph.

4.2.3 Comparing Spread in Lattice and Random Graphs

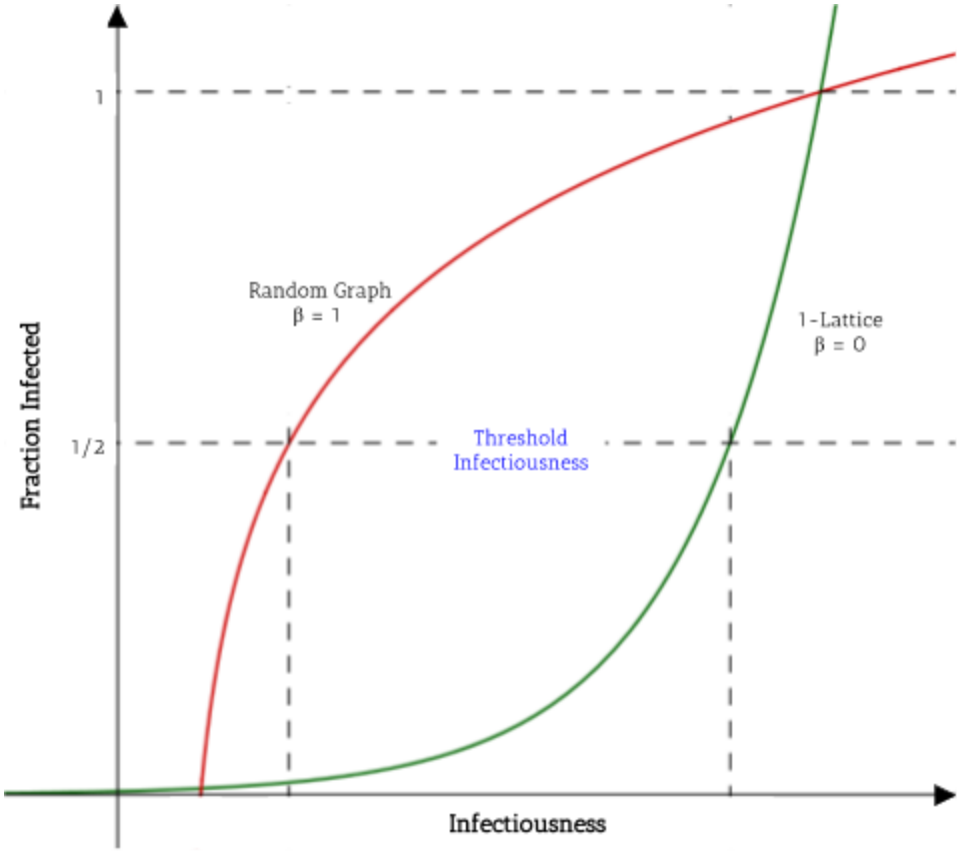


Fig 6. Chart showing how network structure determines the effectiveness of a disease's infectiousness

Fig 6. shows the infection curves with respect to disease infectiousness for spread in a random graph and a 1-lattice. In lattice graphs, reproduction rate is difficult to quantify, so we will speak in terms of infectiousness (our β′ term from the SIR model). Remember that infectiousness (along with death rate) is used to determine reproduction rate, and hence the two are highly correlated – especially given that death rate is not dependent on network structure to near the same degree as infectiousness. We consider threshold infectiousness to be the point at which half the population is infected.

From *Fig 6.* we can determine that in a random graph, threshold infectiousness is low, in that it does not take a very infectious disease to infect half of the population. This is what we hypothesised, given that the creation of shortcuts allows the disease to cross previously long distances. The opposite is true of a lattice graph, where the threshold infectiousness is high and it takes a very infectious disease to infect half the population. This is also intuitive given our discussion of how reproduction rate is slow to increase since infectives already have a high number of infected connections.

This idea gives us true insight into how shortcuts can determine how much of the population will become infected by a disease. Remember a 1-lattice has no shortcuts, but the minute we perform beta rewiring of even a very low β value and create even 1 shortcut, infectiousness becomes dramatically more important. As such, a small world network can have a threshold infectiousness equivalent to that of a random graph and still be, in reality, far from a random network (consider again *Fig 1.*). This is dangerous, since the combination of various modern technologies bring about these random connections with ease, creating a network with the worst case threshold infectiousness.

4.3 Controlling the Spread of Disease

The combination of understanding the SIR model, and disease spread in a small world, allows world health organisations to effectively combat epidemics through targeted aid to places of high clustering, and minimising connections and susceptibility. These insights provide us with the necessary information to efficiently allocate resources, particularly doctors and medication, so that epidemics can be controlled, (and prevented), in a timely manner. This will not only reduce sickness and deaths, but will also provide a level of understanding that can bring comfort in knowing what could come should a particular disease break out.

Furthermore, we have seen thoroughly that in a small-world network the key to explosive growth of a disease is a network's shortcuts. Diseases do not spread

very effectively on lattices, and although the small world networks exhibit some important features of random graphs, they still share with lattices the property that locally, most contacts are highly clustered. So *locally,* the growth of a disease behaves very much like it does on a lattice: infected individuals interact mostly with other already infected individuals, preventing the disease from spreading rapidly into the susceptible population. Only when the disease reaches a shortcut does it start to display the worst-case threshold infectiousness, and random mixing behavior.

There is hope for us in that, unlike epidemics on a random graph, epidemics in a small-world network have to survive first through a slow-growth phase (see the 1-Lattice growth curve *Fig 6.*), during which they are most vulnerable, and the lower the density of shortcuts, the longer this slow-growth phase will last. However, since diseases blindly probe networks for shortcuts, they will eventually find and pursue every shortcut if not stopped somehow. The best way to prevent epidemics, then, is to identify and reduce shortcuts between neighbourhoods on the large scale of connections and keep our network as lattice-like as possible.

5 Modeling the spread of computer viruses

So far, as we have been discussing spread in a small world network, we have been referring to the spread of infectious diseases. Now we wish to look at the spread of the electronic disease known as a computer virus. Computer viruses started off life in a more physical sense. Those powerful enough to do damage to a computer were difficult to send over the web, and needed to be transferred hard copy in the form of a floppy disk. Now, however, computer viruses' stomping ground is our email servers. And so, while a person is obviously not their computer, it is fair to consider that you have an email address book which can be considered your virtual neighborhood. At this point, we ought to consider that email address books may be more highly clustered than a person's actual list of connections, since

email is biased towards communication within business and education communities. However, our email servers still make up a small world network.

The SIR model is less relevant for computer viruses: there is not a period of infection; there is no equivalent to probability of infection once connected, other than the factor of human error; we are less concerned about the loss of a computer than the loss of a human life, and so death rate and our count for the removed population is far less applicable. However, a consideration of network vulnerability and robustness is highly pertinent.

5.1 The Melissa Virus and Microsoft's Oversight

The first computer virus to obtain widespread public recognition, known as the Melissa virus (1999), exposed the vulnerability of the email server network, giving Microsoft a true nightmare. The Melissa virus, once opened, is able to access the user's full contact list, and send itself to the first 50 people. The virus is blown up to a massive scale as it spans across the network. It is important to note, however, that the Melissa virus was at large during the prime time of email platform Microsoft Outlook. It soon became known that the Melissa virus could only activate and spread when opened in Outlook. This shows the true oversight on Microsoft's part: creating universal software gives everybody the *exact* same weaknesses, and hence people are susceptible to the same computer viruses. All it takes is one person to identify a single loophole in the software, and the whole network can become infected.

5.2 Controlling Spread of Computer Viruses

So, how can we create a more robust online network? By using different softwares. Today, Apple's iPhone mail app and Gmail make up more than 50% of the email client market. While this is an improvement from from 1999, we can still go further to increase the robustness of our online network. This should not only go for email client software, but for all computer softwares, since viruses are not only

transferred through emails. Currently, Chrome 50.6 and Internet Explorer 11.0 together make up 49% of the browser market share, and Windows 7 alone makes up for 49% of the operating system market share, (and together all Windows operating systems make up 85% of the market share). These are not comforting numbers[2].

While companies are updating their software regularly to combat viruses, further diversity in the network would be a more thorough approach to decreasing vulnerability. We are likely shying away from this solution with tech companies slowly gaining control of the market in general – it is not really in their interest for us to vary our product use. Furthermore, it seems that with each software update and each new piece of technology, brands are advertising higher compatibility. This appears to make our lives easier by increasing efficiency, when in fact, the absence of incompatibility and diversity is only making us more vulnerable to infection.

6 Conclusions and Thoughts for Future Development

Applying graph theory to modeling the spread of disease and computer viruses in a small world world gives insights into what elements of our network are detrimental and beneficial to promoting or slowing spread. We can take action to combat these negative aspects by changing the inherent nature of our network, in reducing shortcuts and implementing diversity, or by using target activity to repair areas where shortcuts and uniformity are most numerous and most prominent.

This theory or using graph theory for the small world problem could be made further applicable with the derivation of concrete equations that give accurate numerical evaluations of the relationships between reproduction rate, infectiousness, clustering coefficient, and mean path length. With the current model, we only have general correlations, speculations and approximations; there are few

[2] Market share stats retrieved from: emailclientmarketshare.com and netmarketshare.com

direct relationships that we can quantify. These innovations would increase the accuracy of our predictions for spread behavior, allowing us to combat diseases and computer viruses more effectively. Overall, however, the knowledge that we have developed so far has shown relationships relating to the nature of our network that should absolutely affect the way in which we handle epidemics.

Citations:

Astacio, J., Briere, D., Guillén, M., Martinez, J., Rodrigues, F., & Valenzuela-Campos, N. (1996). Mathematical Models to Study the Outbreaks of Ebola. *Cornell eCommons*. Retrieved from ecommons.cornell.edu/bitstream/handle/1813/31962/BU-1365-M.pdf?sequence=1&isAllowed=y

Duncan J. Watts. (1999). *Small Worlds: The Dynamics of Networks between Order and Randomness*. Princeton University Press.

Email Client Market Share. (2017, April). Retrieved from https://emailclientmarketshare.com/

NET MARKET SHARE. (2017, April). Retrieved from www.netmarketshare.com/browser-market-share.aspx?qprid=2&qpcustomd=0; www.netmarketshare.com/operating-system-market-share.aspx?qprid=10&qpcustomd=0

Northcutt, S. (1999, April 22). What was the Melissa virus and what can we learn from it. Retrieved from www.sans.org/security-resources/idfaq/what-was-the-melissa-virus-and-what-can-we-learn-from-it/5/3#13

Nykamp, D. (2016). Small world networks. Retrieved from mathinsight.org/small_world_network

Richard J. Trudeau. (1993). *Introduction to Graph Theory*. New York: Dover Publications, INC.

Watts, D. J. (2003). Epidemics and Failures. In *Six Degrees* (pp. 163 – 194). W.W. Nortan & Company.

Sarah Dennis
December 16th 2016

Mathematical Models for Fire Spread

Analysing their derivations, limitations, and applications

Abstract

Published in 1972, Richard C. Rothermel designed a mathematical model for predicting the spread and intensity of forest fires. His work finds a solution to the integral equation for the general rate of spread for a fire from the Frandsen Model (1971). Both Frandsen's and Rothermel's models rely on the work of Fons, who applied the principle of conservation of energy to fire modeling, whereby knowing what fuel a fire is consuming, you can estimate the fire's intensity, and thus its rate of spread. Rothermel's model finds experimental values for fuel parameters in order to solve Frandsen's equation, incorporating considerations for wind and slope, and finally, providing adaptations for heterogeneous fuels.

Overall, Rothermel provides an accurate estimation of fire behavior, especially for a model of its time. The Rothermel Model is easy to work with, and is still widely used today. It does have several limitations, such as its use of empirical measurements, and its presumption that fuel is continuous, burns uniformly, and that only moisture will allow for extinguishment. In his 1972 publication, Rothermel claims that his model is only suitable for predictions and precautionary measures. Today, computer programs have been developed that enable us to use his model for live fires. We are quickly moving past this, however, and there have been recent requests for a new model more adapted to the technology we have available to us in the twenty first century.

Being an analysis of Rothermel's A Mathematical Model for Predicting Fire Spread in Wildland Fuels, this project will work through the derivation of the equations presented in Rothermel's guide, and then resume a study of the accuracy of his model, examining its functionality and efficacy given our present technologies.

1. Introduction

A mathematical model predicting wildfire spread is an important asset to all those working in fire fighting and fire safety. Fire's innate unpredictability is a large part of the destruction that too often transpires. In 1960, the United States Forest Service recruited engineers and mathematicians, notably Harry Gisborne, Jack Barrows, Richard Rothermel, Hal Anderson, and Bill Frandsen, to study fire behavior at the new Missoula Lab so as to improve fire safety and fire prevention practices [1]. Together, the equations of Frandsen and Rothermel provide a thorough approximation of the rate of fire spread, allowing us to prepare for and combat fires more effectively.

2. Frandsen's model for fire spread

2.1 Introduction to Frandsen's Model

Bill Frandsen's equation for rate of fire spread, published in 1971, presents a "ratio between the heat flux received from the source in the numerator, and the heat required for ignition by the potential fuel in the denominator"(Rothermel, 1972). There is an essential logic to Frandsen's equation when we consider the principle of

[1] Sullivan, A. L. (2008). The Rothermel Fire-Spread Model: Still Running Like a Champ. *Fire Science*, , (2)

conservation of energy. The total energy produced in a combustion reaction cannot exceed the energy available in the fuel source. In other words, the growth rate of the fire is limited by the energy of available fuel. As such, parameters limiting fire spread are on the denominator of Frandsen's equation, and parameters facilitating spread fall in the numerator of the equation.

2.2 FRANDSEN'S EQUATION

Frandsen's model for rate of fire spread,

$$R = \frac{I_{xig} + \int_{-\infty}^{0} (\frac{\delta I_z}{\delta z})_{z_c} dx}{\rho_{be} Q_{ig}}$$ [2], $[ft/min]$

shows the rate of spread as equal to the propagating flux, as in the horizontal flux plus the integral from negative infinity to the interface of the slope of vertical reaction intensity, divided by the product of effective bulk density and heat for pre-ignition.

Frandsen uses x to denote the horizontal direction, and z to denote the vertical direction. Propagating flux, I_p, refers to the total amount of heat available at the fire front to provide forward movement. It equates to the sum of all variables encouraging fire spread,

$$I_p = I_{xig} + \int_{-\infty}^{0} (\frac{\delta I_z}{\delta z})_{z_c} dx$$ [3], $[B.t.u/ft^2 min]$

and forms the numerator of Frandsen's equation. Horizontal heat flux, I_{xig} [$B.t.u/ft^2 min$], represents the amount of heat energy absorbed by a unit volume of fuel at time of ignition. z_c [ft] refers to the fuel bed depth, considered constant by

[2] Rothermel (1972), Equation 1
[3] Rothermel (1972), Equation 5

Frandsen. Evaluated at z_c, the slope of vertical intensity, $(\frac{\delta I_z}{\delta z})$ [$B.t.u/ft^3 min$], is integrated from $x = -\infty$ to $x = 0$ to account for a fixed reaction zone, whereby the unit volume moves towards the interface at $x = 0$ whereupon the fuel is ignited. Effective bulk density, ρ_{be} [lb/ft^3], refers to the total amount of fuel raised to ignition ahead of a fire per unit volume fuel of the fuel bed. Finally, the heat for pre-ignition, Q_{ig} [$B.t.u/lb$], represents the heat required to bring one unit weight of fuel to ignition.

2.3 EFFECTIVENESS AND SOLUTIONS FOR IMPROVEMENT

Ineffectiveness in Frandsen's model spurs from the difficulty one experiences in attempting to quantify each of the components. One of Rothermel's aims in adapting Frandsen's equation is to simplify the parameters to ones that can be easily calculated. Furthermore, Frandsen's model assumes homogenous fuel, with no slope nor wind to add to propagating flux. These are rare and idealistic circumstances. As such, the model has only minor applicability. This model is applicable to farming scenarios or similar, where fuel is homogenous and the ground is flat, however, we have still not accounted for the effect of wind. Following the additions to Frandsen's model, we will look at Rothermel's addition of entirely new variables that account for wind and slope.

3. Rothermel's adaptations to Frandsen's model

In order to simplify and solve Frandsen's rate of spread equation, Rothermel seeks assistful relationships between parameters in R and variables we are able to measure with ease. To this end, he introduces several new concepts relating to fire

behaviour, and redefines several concepts we have recently become familiar with. The primary elements of concern in Rothermel's adaptation of Frandsen's model are effective heating number, heat for pre-ignition, propagating flux, reaction velocity and reaction intensity.

3.1 CALCULATING EFFECTIVE HEATING NUMBER

The effective heating number, ξ, is a ratio of the effective bulk density to the actual bulk density:

$$\xi = \frac{\rho_{be}}{\rho_b} \quad [4]$$

Through experimental calculation, it can be found that

$$\xi = e^{(-138/\sigma)} \quad [5],$$

where σ is the fuel particle surface-area-to-volume ratio. This ratio can be found using

$$\sigma = \frac{4}{d} \quad [6] \; [ft^{-1}]$$

where d is the particle's diameter (for circular particles) or edge length (for square particles).

3.2 CALCULATING HEAT FOR PRE-IGNITION

Calculating heat for pre-ignition relies primarily on M_f, the ratio of fuel moisture to oven-dry weight, and T_{ig}, the ignition temperature of the ground. Heat for pre-ignition also considers the specific heat of dry wood, the specific heat of water, the temperature range of boiling for water, and the latent heat of vaporization for

[4] Rothermel (1972), Equation 3
[5] Rothermel (1972), Equation 14
[6] Rothermel (1972), Equation 32

moisture. Assuming a temperature ignition range for the ground of 20°C to 320°C, and a boiling temperature of water of 100°C, the preignition heat can be written as

$$Q_{ig} = 250 + (1116)M_f \quad [7].$$

3.3 CALCULATING PROPAGATING FLUX

Substitute the effective heating number and propagating flux,

$$\xi = \frac{\rho_{be}}{\rho_b} \qquad \text{and} \qquad I_p = I_{xig} + \int_{-\infty}^{0} (\frac{\delta I_z}{\delta z})_{z_c} dx,$$

Into Frandsen's rate of spread equation under the conditions of no wind where

$$I_p = (I_p)_0 \quad \text{and} \quad R = R_0,$$

to achieve

$$R_0 = \frac{(I_p)_0}{\xi \rho_b Q_{ig}}.$$

Rewritten in terms of propagating flux,

$$(I_p)_0 = R_0 \rho_b \xi Q_{ig} \quad [8],$$

the equation clearly shows that propagating flux for a no wind fire is equal to the rate of spread for that fire, multiplied by its effective bulk density, multiplied by the heat for pre-ignition.

3.4 CALCULATING REACTION INTENSITY PART I

Changes in energy at the fire front are the result of a chemical combustion reaction in organic matter. We can use "the rate of change of this organic matter from a solid to a gas" as "a good approximation of the subsequent heat release rate of the

[7] Rothermel (1972), Equation 12
[8] Rothermel (1972), Equation 6

fire"(Rothermel, 1972). Reaction intensity, I_R, represents the heat release rate per unit area at the fire front. As a rate, we can represent reaction intensity as a differential equation,

$$I_R = -\frac{dw}{dt}h \text{ [9]}, [B.t.u/ft^2 min]$$

where $\frac{dw}{dt}$ is mass lost per unit area with respect to time at the fire front and h is the heat content of the fuel. By the chain rule, reaction intensity can be expressed as a product of derivatives,

$$I_R = -(\frac{dw}{dx})(\frac{dx}{dt})h \text{ [10]}$$

where $\frac{dx}{dt}$ represents R the rate of spread of the fire in horizontal distance with respect to time. Thus,

$$I_R = -(\frac{dw}{dx})Rh.$$

We can solve this as a separable integral equation in which is evaluated over the reaction zone depth, and is evaluated over the limits of the loading in the reaction zone. Fuel loading refers to the amount of fuel available for combustion. The separable integral takes the form

$$I_R \int_0^D dx = -Rh \int_{W_n}^{W_r} dw \text{ [11]}$$

and simplifies to

[9] Rothermel (1972), Equation 7
[10] Rothermel (1972), Equation 15
[11] Rothermel (1972), Equation 17

$$I_R D = Rh(w_n - w_r)$$ [12]

where w_n is the net initial fuel loading, and w_r is the residue loading immediately after the passing of reaction zone.

A term for reaction time can be introduced, equivalent to the time taken for the fire front to travel the depth of the reaction zone:

$$\tau_R = \frac{D}{R}$$ [13].

Inputting reacting time into

$$I_R D = Rh(w_n - w_r)$$

Gives reaction intensity in terms of reaction time:

$$I_R = \frac{h(w_n - w_r)}{\tau_R}$$ [14].

Notice that the maximum reaction intensity occurs when residue loading is zero:

$$I_{R\max} = \frac{hw_n}{\tau_R}$$ [15].

The efficiency of the reaction zone can now be found as a ratio between actual reaction intensity and maximum reaction intensity:

$$n_\delta = \frac{I_R}{I_{R\max}}$$ [16] or $$n_\delta = \frac{h(w_n - w_r)}{\tau_R} \cdot \frac{\tau_R}{hw_n} = \frac{w_n - w_r}{w_n}.$$

From this we can obtain that

---

[12] Rothermel (1972), Equation 18
[13] Rothermel (1972), Equation 19
[14] Rothermel (1972), Equation 20
[15] Rothermel (1972), Equation 21
[16] Rothermel (1972), Equation 22

$$w_n - w_r = w_n \cdot n_\delta$$

Inputting this into

$$I_R = \frac{h(w_n - w_r)}{\tau_R}$$

gives

$$I_R = \frac{w_n \cdot n_\delta \cdot h}{\tau_R}$$ [17].

This represents reaction intensity in terms of parameters we can easily measure when given reaction velocity.

3.5 CALCULATING REACTION VELOCITY

Reaction velocity indicates the rate of fuel consumption, and the degree to which fuel is consumed, in other words, it represents "the dynamic character of the fire"(Rothermel, 1972). Reaction velocity can be represented as a ratio between the reaction zone efficiency, $n\delta$, and the reaction time, τ_R:

$$\Gamma = \frac{n_\delta}{\tau_R}$$ [18] [min^{-1}].

A fuel bed's moisture content, mineral content, particle size, and bulk density all affect the reaction velocity. As such, we write that

$$\Gamma = \Gamma' n_M n_s$$ [19]

where Γ' [min^{-1}], represents potential reaction velocity, n_M is the moisture damping coefficient and n_s is the mineral damping coefficient. These values require experimental evaluation.

---

[17] Rothermel (1972), Equation 23
[18] Rothermel (1972), Equation 25
[19] Rothermel (1972), Equation 26

The moisture damping coefficient is defined as a ratio of actual reaction intensity to maximum reaction intensity given that the moisture content of the fuel is zero:

$$n_M = \frac{I_R}{I_{R\max}} \quad [20].$$

Experiments by Rothermel's co-worker, Anderson (1969), provide a value for the moisture damping coefficient:

$$n_M = 1 - 2.59\frac{M_f}{M_x} + 5.11(\frac{M_f}{M_x})^2 - 3.52(\frac{M_f}{M_x})^3 \quad [21]$$

where M_f represents the present fuel moisture, and M_x denotes the moisture level at which the fire will not spread.

The mineral damping coefficient was evaluated by Rothermel's coworker Philpot in 1968. The experiments determined that

$$n_s = 0.174(S_e)^{-0.19} \quad [22]$$

where S_e is the effective mineral content.

After gathering this experimental data, we can write an expression for reaction velocity,

$$\Gamma' = \Gamma'_{\max}(\beta/\beta_{op})^A exp[A(1-\beta/\beta_{op})] \quad [23]$$

where

$$\Gamma'_{\max} = \frac{\sigma^{1.5}}{495 + 0.0594\sigma^{1.5}} \quad [24],$$

[20] Rothermel (1972), Equation 28
[21] Rothermel (1972), Equation 29
[22] Rothermel (1972), Equation 30
[23] Rothermel (1972), Equation 38
[24] Rothermel (1972), Equation 36

$$\beta_{op} = 3.348\sigma^{-0.8189} \quad [25],$$

$$A = \frac{1}{4.77\sigma^{0.1} - 7.27} \quad [26],$$

σ is the particle surface-area-to-volume ratio, β represents the fuel bed compactness, and β_{op} is the 'optimal' level. Fuel bed compactness represents a ratio between the fuel bulk density, ρ_b [lb/ft^3], and the fuel particle density, ρ_p [lb/ft^3]:

$$\beta = \frac{\rho_b}{\rho_p} \quad [27]$$

Note that fuel burns best when $\beta/\beta_{op} = 1$.

3.6 CALCULATING REACTION INTENSITY PART II.

Having calculated reaction velocity, we can substitute the two equations,

$$\Gamma = \frac{n_\delta}{\tau_R} \qquad \text{and} \qquad \Gamma = \Gamma' n_M n_s$$

into

$$I_R = \frac{w_n \cdot n_\delta \cdot h}{\tau_R} \quad [28],$$

such that

$$I_R = w_n \cdot h \cdot \Gamma' n_M n_s \quad [29]$$

reaction intensity is now in terms of measurable parameters.

3.7 THE RATIO BETWEEN PROPAGATING FLUX AND REACTION INTENSITY

We now introduce the ratio between propagating flux and reaction intensity:

---

[25] Rothermel (1972), Equation 37
[26] Rothermel (1972), Equation 39
[27] Rothermel (1972), Equation 31
[28] Rothermel (1972), Equation 23
[29] Rothermel (1972), Equation 27

$$\zeta = \frac{(I_p)_0}{I_R}$$ [30] or $$(I_p)_0 = \zeta I_R$$

By plotting ζ as a function of different fuel bed compactness levels, we can see the correlation:

$$\zeta = (192 + 0.259\sigma)^{-1} exp[(0.792 + 0.681\sigma^{0.5})(\beta + 0.1)]$$ [31].

This formula shows the ratio of propagating flux to reaction intensity in terms of the same variables as reaction velocity. This will simplify applications of the model by reducing the environment-specific parameters one needs to measure.

3.8 Rothermel's Final Adaptation of Frandsen's Model

By returning to the rate of spread for a no-wind fire, which we defined as,

$$R_0 = \frac{(I_p)_0}{\xi \rho_b Q_{ig}}$$

and inputting

$$(I_p)_0 = \zeta I_R$$

Into this relationship, we see that

$$R_0 = \frac{I_R \zeta}{\rho_b \xi Q_{ig}}$$ [32].

This is Rothermel's equation for the rate of spread of a fire when there is no wind and no slope. Each element of this equation has a method for calculation that, either using raw data calculated in the field of application, or a formula that can transform this data

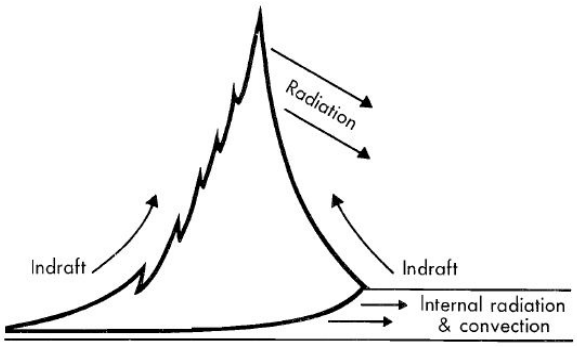
[30] Rothermel (1972), Equation 41
[31] Rothermel (1972), Equation 42
[32] Rothermel (1972), Equation 43

into an applicable variable. This formula directly correlates to data samples gathered by Rothermel, confirming the accuracy of his additions to Frandsen's work.

4. Rothermel's Rate of Spread Equation

4.1 Incorporating Wind and Slope

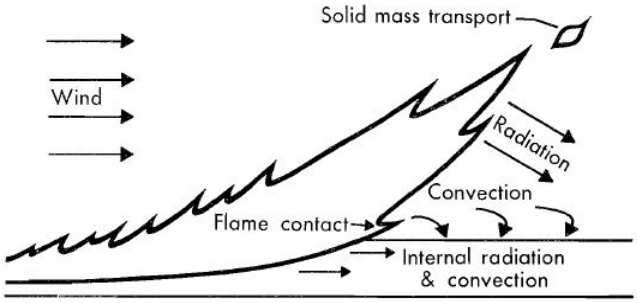
Rothermel's true work begins with his incorporation of wind and slope into the rate of fire spread model. This practice increases the accuracy and effectiveness of the mathematical model beyond the scope of the accuracy of Frandsen's model. By observing the basic principles of fire behavior, it is intuitive how we should incorporate wind and slope into the equation. Heat inherently travels upwards. Thus, in a scenario with level ground and no wind, flames are traveling vertically and the fire moves forwards relatively slowly, due to its heat moving primarily away from the fuel.



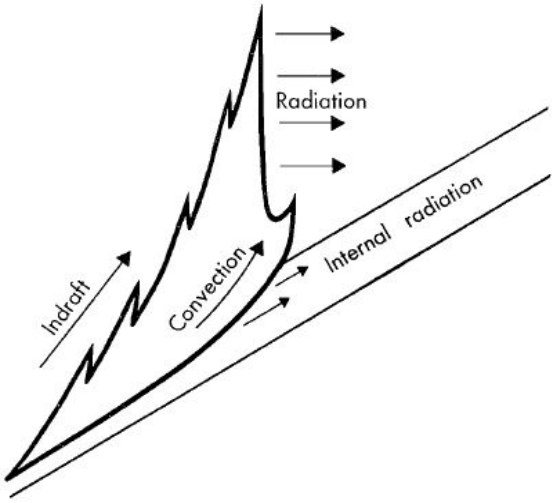
Similarly consider the scenario of a fire on flat ground where wind is blowing parallel to the surface. The flames, previously travelling straight upwards, are now being forced to follow the direction of the wind, bringing them in greater contact with

[33] Rothermel (1972), Fig 2.

the ground. The fuel is receiving a larger amount of heat from the fire, causing the fire to spread more quickly.



By adding an up-slope, we are affecting the scenario in essentially the same manner as with wind. An upslope brings the fuel closer to the flames, providing the ground with more heat and allowing the fire to travel faster. In summary, wind and slope "change the propagating heat flux by exposing the potential fuel to additional convective and radiant heat"(6).



[34] Rothermel (1972), Fig 3.
[35] Rothermel (1972), Fig 4.

4.2 Rothermel's Final Rate of Spread Equation

Knowing that wind and slope increase propagating flux, we obtain that the propagating flux for a wind and slope affected fire, I_P, ought to be represented as

$$I_p = (I_p)_0(1 + \phi_w + \phi_s)$$[36]

where ϕ_w and ϕ_s represent the wind and slope coefficients (respectively). Since propagating flux resides on the numerator of (a variation of) Rothermel's rate of spread equation,

$$R_0 = \frac{(I_p)_0}{\xi \rho_b Q_{ig}},$$

we can easily adapt this formula to consider the a wind and slope affected fire:

$$R = \frac{(I_p)_0(1 + \phi_w + \phi_s)}{\rho_b \xi Q_{ig}}$$[37].

Furthermore, we can incorporate

$$(I_p)_0 = \zeta I_R,$$

so that the equation for rate of spread,

$$R = \frac{I_R \zeta (1 + \phi_w + \phi_s)}{\rho_b \xi Q_{ig}}$$[38]

continues to work with those variables requiring the fewest environment-specific calculations. This is Rothermel's final equation for rate of spread. We now only need to

[36] Rothermel (1972), Equation 9
[37] Rothermel (1972), Equation 10
[38] Rothermel (1972), Equation 52

evaluate wind and slope experimentally before we can determine the rate of spread for a fire.

4.3 CALCULATING WIND COEFFICIENT FOR PROPAGATING FLUX

To quantify the wind coefficient –the constant that increases the propagating flux of a no wind fire – contributing in final to the propagating flux of a wind-affected fire, take

$$I_p = (I_p)_0(1 + \phi_w + \phi_s)$$

and let $\phi_s = 0$. Then, the coefficient for propagating flux from wind,

$$\phi_w = \frac{I_p}{(I_p)_0} - 1 \quad [39],$$

is a ratio between the actual propagating flux and the propagating flux given no wind, minus 1. Now, take the equation

$$(I_p)_0 = R_0\rho_b\xi Q_{ig},$$

let the fuel parameters be constant, and consider the resultant relationship

$$\phi_w = \frac{I_p}{(I_p)_0} - 1.$$

See that this can be equated to

$$\phi_w = \frac{R_w}{R_0} - 1 \quad [40],$$

representing the rate of spread in the presence of a head wind. Rothermel's experiments provide the real value for the wind coefficient to be

[39] Rothermel (1972), Equation 44
[40] Rothermel (1972), Equation 45

$$\phi_w = CU^B(\beta/\beta_{op})^{-E} \quad [41]$$

where

$$C = 7.47\exp(-0.133\sigma^{0.55}) \quad [42],$$

$$B = 0.02526\sigma^{0.54} \quad [43],$$

$$E = 0.715\exp(-3.59 * 10^{-4}\sigma) \quad [44],$$

U is the wind velocity [*ft/min*], σ is the particle surface-area-to-volume ratio, β is fuel bed compactness and β_{op} is the 'optimal' level of fuel compactness.

4.4 CALCULATING SLOPE COEFFICIENT FOR PROPAGATING FLUX

Likewise for the slope coefficient as with the wind coefficient,

$$\phi_s = \frac{R_s}{R_0} - 1 \quad [45],$$

Gives the rate of spread up a slope. Rothermel's experiments approximate the slope coefficient as

$$\phi_s = 5.275\beta^{-0.3}(\tan\theta)^2 \quad [46]$$

where $\tan\theta$ is the slope of the fuel bed and β is fuel bed compactness. Presumably we could calculate the slope of the fuel bed using other trigonometric identities based off of what information we know of the environment. Note that as slope increases, ϕ_s increases, and so the rate of spread will increase too.

---

[41] Rothermel (1972), Equation 47
[42] Rothermel (1972), Equation 48
[43] Rothermel (1972), Equation 49
[44] Rothermel (1972), Equation 50
[45] Rothermel (1972), Equation 46
[46] Rothermel (1972), Equation 51

4.5 Approximate Value Input

We can now test Rothermel's final equation using approximate values [47] – those that are common or average for a standard forest fire:

Reaction intensity, $I_R = 2555 B.t.u/ft^2 min$

Propagating flux ratio, $\zeta = 0.8$

Wind coefficient, $\phi_w = 3$

Slope coefficient, $\phi_s = 3$

Fuel bulk density $\rho_b = 32 lb/ft^3$

Effective heating number $\xi = 0.2$

Heat for pre-ignition $Q_{ig} = 200 B.t.u/lb$

Then the rate of spread,

$$R = \frac{2555 * 0.8(1+3+3)}{32 * 2 * 200} = 11.2 ft/min.$$

This is a slow rate of spread, but often forest fires are over moist ground, and we considered slope and wind to be present but minimal. This is a reasonable value for us to obtain from Rothermel's equation.

5. Applications and Functionality of rothermel's model

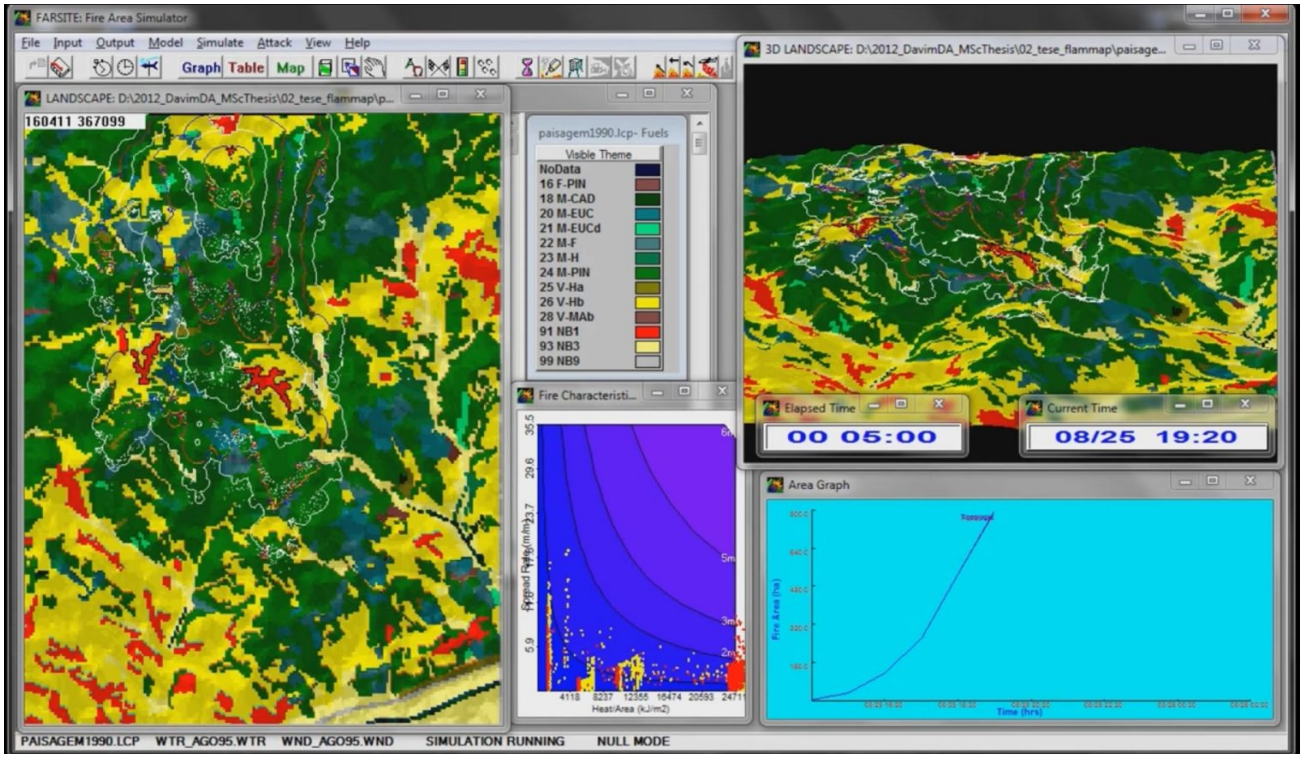
In 1972, Rothermel professed that his equation is only suitable for two scenarios: (1) hypothetical fires, as in fuel appraisal, fire planning, and fire training; and (2) possible fires, as in fire danger ratings and pre-suppression plannings. He claims that

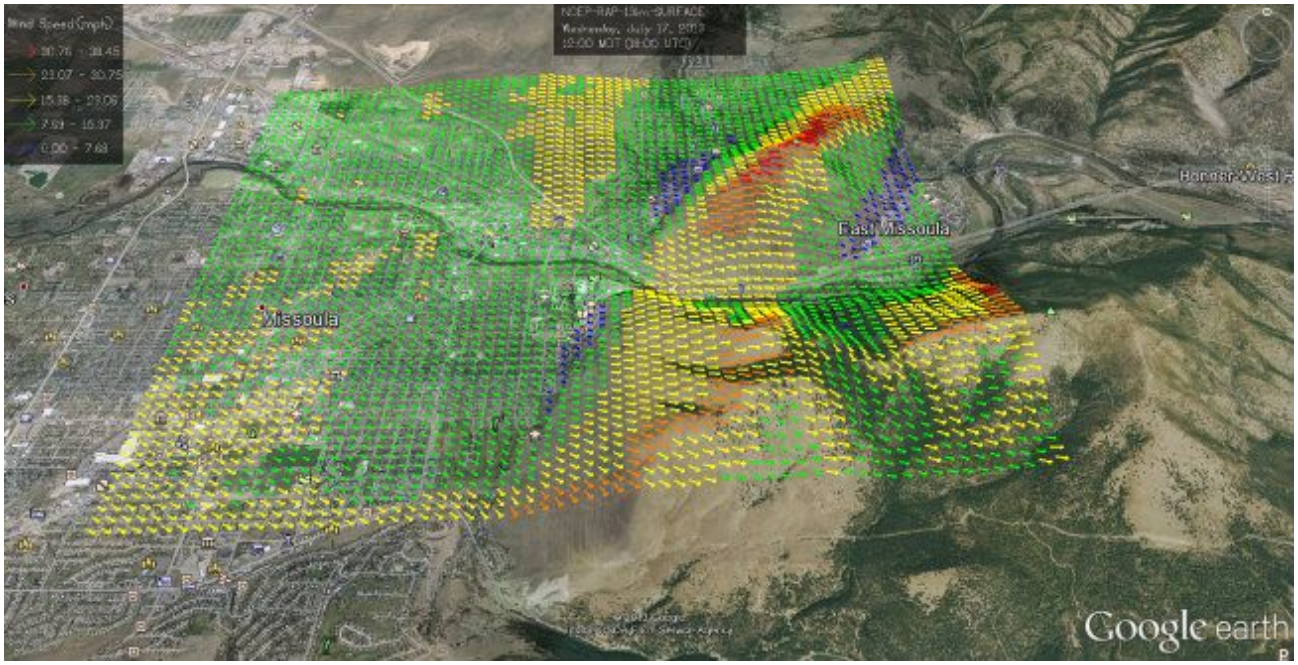
[47] Sugihara, N. G. (2006)

"forecasting the behavior of existing wildfires will require a greater degree of sophistication than this model and our knowledge of fuels will permit at this time"(Rothermel, 1972). In 1983, Rothermel published a guide to facilitate calculations of rate of fire spread. This document includes detailed step-by-step methods for measuring the parameters of an environment, and several tables of estimated values for these parameters that one could employ if found without the resources to take quantifiable observations themselves. Rothermel's guide significantly broadens the spectrum of his model's applicability by allowing it to be applied with greater ease and less time, and by those with less mathematical or scientific background.

Today, we have come further still. There now exist numerous 'Decision Support Systems' for fire managers that simplify, mimic, and add to Rothermel's model. These programs include BEHAVE, FARSITE, FlamMap, FireFamily Plus, Rare Event Risk Assessment Process, WindNinja, FireStem, and Wildland Fire Assessment System.[48] Some systems simply calculate the raw data from Rothermel's model, some give more accurate measures of environment-specific parameters, while others are able to provide 3D visual representations of fire spread predictions. Each system has a unique value to those wishing to pursue fire analysis and forecast.

[48] Wells, G. (2008). The Rothermel Fire-Spread Model: Still Running Like a Champ. , (2).





[49] David A. Davim. (2012). .

[50] Mark Vosburgh. (2013, July 18). Missoula Scientists Study Wind Flow and the Affects on Firelines.

6. Limitations of Rothermel's Model & areas for improvement

Rothermel's fire spread model has several limitations. It presumes that fuel is continuous, burns uniformly, and that only moisture will allow for extinguishment – when often this is not entirely true. Comments by Mark Finney, one of Rothermel's successors and a director of FARSITE, express the need of a new model to address our contemporary concerns: "New demands are being placed on those old models. We're asking them to do things they were not designed to do, to answer questions that didn't have a practical context then" (Wells, 2008). As with most things, it is essential to persist with innovation to maintain effectiveness, accuracy, and accessibility.

One example Finney considers of high priority for development is a model that can be applied to the burning off process – where forest fires are intentionally started, in a controlled manner, so as to prevent those that could occur in the future in less than ideal conditions. The ability to accurately predict fire behavior could significantly improve the safety, accuracy, and accessibility of this process, working to make forests safer in the long term. Such a model requires us to "[look] more deeply into the three modes of heat transfer—conduction, convection, and radiation—in an attempt to understand the actual mechanisms of fire spread" (Wells, 2008).

It is essential that, in the process of innovation, we do not lose sight of the importance of accessibility of the mathematical model. It would be improper for a new method to be unmanageably complex or disproportionately expensive as a result of technological reliance, for then its applicability as a universal tool for fire prediction is less than that of the Rothermel Model. Forest fires are of greatest threat in rural areas

where both personnel and monetary resources, as needed to operate a complex model, are highly limited. This leaves us with a lofty, but conceivably achievable goal for the future: to fully understand the mechanics of fire and devise a model that is accessible to all those wishing to understand fire behavior. For fire, both as a danger and as a refugee, is present in so many fields.

Citations

Ali Karouni, Bassam Daya, Samia Bahlak, Pierre Chauvet. (2014). A Simplified Mathematical Model for Fire Spread Predictions in Wildland Fires Combining between Models of Anderson and Rothermel. , (3). Retrieved from http://www.ijmo.org/papers/372-A1023.pdf

David A. Davim. (2012). . Retrieved from https://www.youtube.com/watch?v=gLx9GOPALpE

Jo, The Applied and Industrial Mathematics Research Group. (2013, July 11). Maths of Planet Earth | Limitless Applications. Retrieved from http://mathsofplanetearth.org.au/the-mathematics-of-fire-predicting-the-growth-of-bushfires/

Keane, R. E. (2014). . Springer.

Mark Vosburgh. (2013, July 18). Missoula Scientists Study Wind Flow and the Affects on Firelines. . Retrieved from http://www.makeitmissoula.com/2013/07/missoula-scientists-study-wind-flow-and-the-affects-on-firelines/

Sugihara, N. G. (2006). . University of California Press.

Richard. C. Rothermel. (1972). A Mathematical Model For Predicting Fire Spread in Wildland Fuels. Intermountain Forest and Range Experiment Station. Retrieved from http://www.fs.fed.us/rm/pubs_int/int_rp115.pdf

Richard C. Rothermel. (1983, June). How to Predict the Spread and Intensity of Forest and Range Fires. US Dept. Agriculture; Forest Service. Retrieved from http://www.fs.fed.us/rm/pubs_int/int_gtr143.pdf

Utah State University. (2008). Fuels Classification. Retrieved October 17, 2016, from http://ocw.usu.edu/Forest__Range__and_Wildlife_Sciences/Wildland_Fire_Management_and_Planning/Unit_2__Fuels_Classification_2.html

Weise, D., & Gregory. (1997). A Qualitative Comparison of Fire Spread Models Incorporating Wind and Slope Effects. , (2). Retrieved from http://www.fs.fed.us/psw/publications/4403/psw_1997_weise000.pdf

Wells, G. (2008). The Rothermel Fire-Spread Model: Still Running Like a Champ. , (2). Retrieved from https://www.firescience.gov/Digest/FSdigest2.pdf

The Seifert - van Kampen Theorem

Sarah Dennis

December 18, 2019

1 Introduction

A large part of the study of algebraic topology is using algebraic tools such as to describe and understand topological properties. Homomorphisms, groups, order, generators and binary operations are all concepts drawn from algebra. We particularly find this intersection between algebra and topology when studying the fundamental group of surfaces. But to apply these algebraic techniques, we must have a means of translating topological properties into algebraic properties and vice versa. The Seifert - van Kampen theorem is a good example of how we transition between algebra and topology. Essentially this theorem allows us to classify a surface (by fundamental group) that is a union of smaller surfaces. If we can decompose the surface into more simple parts, we can then simplify the problem of classifying this surface. By applying the Seifert - van Kampen theorem, we are able to bypass intricate algebraic manipulations and the transition from topology to algebra is less involved.

The Seifert-van Kampen theorem results from the work of mathematicians Egbert van Kampen and Herbert Seifert. Seifert's contribution to this theorem came in 1931 with his PhD thesis at the University of Dresden: *Konstruktion dreidimensionaler geschlossener Räume* (translation: Construction of three-dimensional closed spaces) [4]. Then in 1933, van Kampen was working at Johns Hopkins University where he published the article: *On the Connection between Fundamental Groups of Some Related Spaces* [7]. This article builds off of Seifert's work, presenting all elements needed to state the full theorem alongside thorough examples of the theorem's applications to surface classification. The full statement of the theorem is not given however, as van Kampen states it would "be more confusing than helpful"[7]. Hence there are a wide range of varying statements of the Seifert - van Kampen theorem. Each statement uses slightly different terminology but the meaning is in general unchanged.

2 The Theorem

2.1 Preliminary definitions

The following definitions are necessary for stating the Seifert - van Kampen theorem.

Definition 1. A **path** in a topological space X is a continuous mapping $\alpha : [0,1] \to X$. When $\alpha(0) = \alpha(1)$ we say α is a **loop**. A space X is **path connected** if given any two points $a, b \in X$ there exists a path $\alpha : [0,1] \to X$ such that $\alpha(0) = a$ and $\alpha(1) = b$.

Definition 2. A **homotopy** between a paths α and β in a space X is a continuous map $H : X \times [0,1] \to X$ such that $H(x,0) = \alpha(x)$ and $H(x,1) = \beta(x)$.

Definition 3. Given a connected space X and a point $x_0 \in X$, the **fundamental group** of X based at x_0 is the group of equivalence classes of loops based at x_0 under the equivalence relation of path homotopy. It is denoted $\pi_1(X, x_0)$. When X is path-connected, the fundamental group of X is independent of base point so we can simply write $\pi_1(X)$.

Definition 4. A **group homomorphism** between groups G and H is a map $\phi : G \to H$ such that

$$\phi(g_1 * g_2) = \phi(g_1) \star \phi(g_2) \tag{1}$$

for $g_1, g_2 \in G$. If ϕ is a bijection, then ϕ is a **group isomorphism**.

Definition 5. Given $B \subset A$, the injection $i : B \to A$ defined by $i(b) = b$ for all $b \in B$ is called the **inclusion map** from B to A.

Definition 6. Let X be a set and let $R \subseteq F(X)$ where $F(X)$ is the free group on X. Let $\langle\langle R \rangle\rangle$ denote the intersection of all normal subgroups of $F(X)$ that contain R. Then $\langle X \mid R \rangle$ gives a **group presentation** for $F(X)/\langle\langle R \rangle\rangle$. [2]

Definition 7. Let G_0, G_1 and G_2 be groups. Let $\phi_1 : G_0 \to G_1$ and $\phi_2 : G_0 \to G_2$ be homomorphisms. Let $\langle X_1 \mid R_1 \rangle$ and $\langle X_2 \mid R_2 \rangle$ be group presentations for G_1 and G_2 where $X_1 \cap X_2 = \emptyset$. Then the **push-out** $G_1 *_{G_0} G_2$ of

$$G_1 \xleftarrow{\phi_1} G_0 \xrightarrow{\phi_2} G_2 \tag{2}$$

is the group with presentation

$$\langle X_1 \cup X_2 \mid R_1 \cup R_2 \cup \{\phi_1(g) = \phi_2(g) : g \in G_0\} \rangle. \text{ [2]} \tag{3}$$

2.2 Statement of the Seifert - van Kampen theorem

Theorem 1. *Let K be a space which is a union of two path-connected open sets K_1 and K_2 where $K_1 \cap K_2$ is also path connected. Let $\langle X_1 \mid R_1 \rangle$ and $\langle X_2 \mid R_2 \rangle$ be group presentations for $\pi_1(K_1, b)$ and $\pi_1(K_2, b)$ respectively with $X_1 \cap X_2 \neq \emptyset$. Take $b \in K_1 \cap K_2$ and let $i_1 : K_1 \cap K_2 \to K_1$ and $i_2 : K_1 \cap K_2 \to K_2$ be the inclusion maps. Then $\pi_1(K, b)$ is isomorphic to the push-out of*

$$\pi_1(K_1, b) \xleftarrow{i_{1*}} \pi_1(K_1 \cap K_2, b) \xrightarrow{i_{2*}} \pi_1(K_2, b), \tag{4}$$

namely,

$$\langle X_1 \cup X_2 \mid R_1 \cup R_2 \cup \{i_{1_*}(g) = i_{2_*}(g) : g \in \pi_1(K_1 \cap K_2, b)\}. \tag{5}$$

The above statement of the theorem is adapted from that of Mark Lackenby [2]. This specific formulation was chosen for its use of push-outs which allows for a precise statement of the group presentation of the space K.

2.3 Examples

Example 1. Consider the connected sum of two tori $T\#T$. This is a path connected space, that is the union of path-connected open sets with path connected intersection $T/D^2 \cap T/D^2 = S^1$. This space then fits the assumptions of the Seifert - van Kampen theorem.

We already know the fundamental group and a group presentation of the fundamental group of the torus

$$\pi_1(T) = \mathbb{Z} \times \mathbb{Z} \simeq \langle a, b | aba^{-1}b^{-1} \rangle$$

and of the circle

$$\pi_1(S^1) = \mathbb{Z} \simeq \langle s | \emptyset \rangle$$

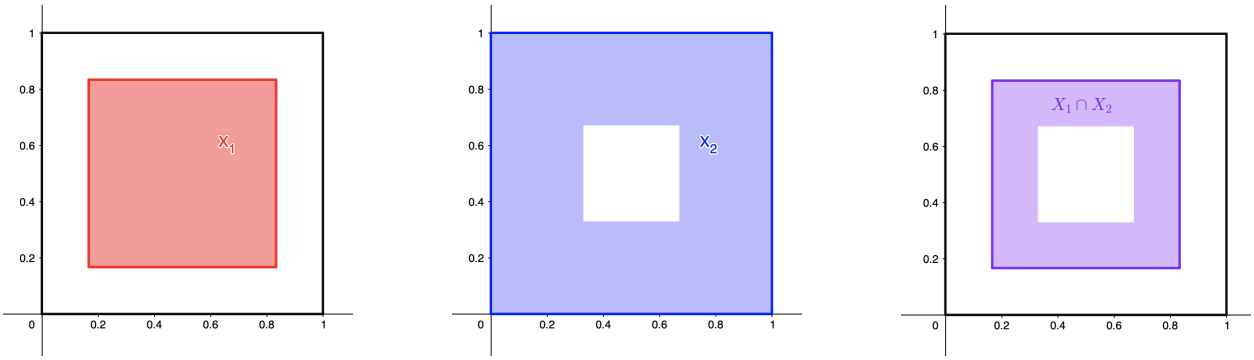
We then need an inclusion map $i : \mathbb{Z} \to \mathbb{Z} \times \mathbb{Z}$, let $i(g) = (g, 0)$. In the statement of the theorem we have i_1 and i_2 but since $K_1 \simeq K_2$ we can use a single inclusion map $i = i_1 = i_2$. This then means that relation $\{i_{1_*}(g) = i_{2_*}(g) : g \in \pi_1(K_1 \cap K_2, b)\}$ is trivial.

Let a, b be generators for the fundamental group of one torus, and let c, d be generators for the other. Then we have

$$\pi_1(T\#T) = \langle a, b, c, d | aba^{-1}b^{-1}, dcd^{-1}c^{-1} \rangle.$$

Now using the fact that $aba^{-1}b^{-1} = dcd^{-1}c^{-1}$ we can write this as a single relation $aba^{-1}b^{-1}cdc^{-1}d^{-1}$ giving

$$\pi_1(T\#T) = \langle a, b, c, d | aba^{-1}b^{-1}cdc^{-1}d^{-1} \rangle.$$



X_1 X_2 $X_1 \cap X_2$

Figure 1: Klein bottle decomposition

Example 2. Consider the Klein bottle K described as the quotient space $[0,1]\times[0,1]\setminus\sim$ where $(s,0)\sim(s,1)$ and $(0,t)\sim(1,1-t)$. Let $X_1=(\frac{1}{6},\frac{5}{6})\times(\frac{1}{6},\frac{5}{6})$ and let $X_2=[0,1]\times[0,1]-(\frac{2}{6},\frac{4}{6})\times(\frac{2}{6},\frac{4}{6})$ as show in figure 1. So $K=X_1\cup X_2$ and $X_1\cap X_2=X_1-(\frac{2}{6},\frac{4}{6})\times(\frac{2}{6},\frac{4}{6})$.

Now, observe that X_1 has trivial fundamental group: $\pi_1(X_1)=\langle a|a\rangle$ and the intersection $X_1\cap X_2$ is isomorphic to S^1 so $\pi_1(X_1\cap X_2)=\langle b|\emptyset\rangle$.

The fundamental group of X_2 is more difficult to determine since the equivalence relation is still relevant. If we retract X_2 to the edges of the unit square, then we can realise X_2 as the wedge of two circles sharing the point $(0,0)=(1,0)=(0,1)=(1,1)$. One circle is points of the form $(0,t)=(1,1-t)$ and the other circle is points of the form $(s,0)=(s,1)$ as we vary $s,t\in[0,1]$. So X_2 has fundamental group $\pi_1(X_2)=\langle c,d|\emptyset\rangle$.

Now, we have the commutative square

$$\begin{array}{ccccc} & & \langle c,d|\emptyset\rangle & & \\ & \overset{\phi_1}{\nearrow} & & \overset{f}{\searrow} & \\ \langle b|\emptyset\rangle & & & & \pi_1(K) \\ & \underset{\phi_2}{\searrow} & & \underset{0}{\nearrow} & \\ & & \langle a|a\rangle & & \end{array} \tag{6}$$

Note that $\phi_1(b)=cdcd^{-1}$ and $\phi_2(b)=a$. Then since we have the zero map from $\langle a|a\rangle$ to $\pi_1(K)$ and this diagram is commutes, we must have $\phi_1(b)=cdcd^{-1}$ as an element of $\ker f$. Furthermore, if there is some other element in $\ker f$ then f would not be surjective. This gives $\pi_1(K)=\langle c,d|cdcd^{-1}\rangle$.

3 Proof

3.1 Definitions

Before we can approach a proof of the Seifert - van Kampen theorem we need several further definitions with related lemmas and theorems.

Definition 8. A **Tietze transformation** is one of the following moves applied to a finite group presentation $\langle x_1,...,x_m|r_1,...,r_n\rangle$. These operations do not change the inherent structure of the group.

(T1) Re-order the generators or relations

(T2) Add or remove the relation e

(T3) Perform an elementary contraction or expansion (inserting or removing $x_ix_i^{-1}$) to a relation r_i

(T4) Insert (or remove) a relation r_i or its inverse into one of the other r_j.

(T5) Add (or remove) a generator x_{m+1} together with a relation $w(x_1, ..., x_m)x_{m+1}^{-1}$ where $w(x_1, ..., x_m)$ denotes any word formed from these letters. [2]

Theorem 2. *The universal property of push-outs [2]. Let $G_1 *_{G_0} G_2$ be the push out of*

$$G_1 \xleftarrow{\phi_1} G_0 \xrightarrow{\phi_2} G_2. \tag{7}$$

Let H be a group and let $\beta_1 : G_1 \to H$ and $\beta_2 : G_2 \to H$ be homomorphisms such that the following diagram commutes:

$$\begin{array}{ccc} G_0 & \xrightarrow{\phi_1} & G_1 \\ \downarrow{\scriptstyle \phi_2} & & \downarrow{\scriptstyle \beta_1} \\ G_2 & \xrightarrow{\beta_2} & G_1 *_{G_0} G_2 \end{array} \tag{8}$$

Then there is a unique homomorphism $\beta : G_1 *_{G_0} G_2 \to H$ such that the following diagram commutes:

$$\begin{array}{ccccc} & & G_1 & & \\ & \swarrow^{\alpha_1} & & \downarrow{\scriptstyle \beta_1} & \\ G_1 *_{G_0} G_2 & & \xrightarrow{\beta} & & H \\ & \nwarrow_{\alpha_2} & & \uparrow{\scriptstyle \beta_2} & \\ & & G_2 & & \end{array} \tag{9}$$

Proof. By definition, the push-out $G_1 *_{G_0} G_2$ has generators $G_1 \cup G_2$. So that the diagram in (9) commutes, define β on these generators by $\beta(x_i) = \beta_i(g_i)$ for $g_i \in G_i$. Now, if β exists, it must be unique.

We check that β is well defined by verifying that $\beta(r) = e_H$ for any relation r in $G_1 *_{G_0} G_2$ and e_H the identity in H. If $r \in G_1$ or $r \in G_2$ then $\beta(r) = e_H$ holds because β_1 and β_2 are homomorphisms. Otherwise, r is of the form $\phi_1(g) = \phi_2(g)$ for some $g \in G_0$. But from the commutativity of the diagram in (8) we have

$$\beta_1(\phi_1(g)) = \beta_2(\phi_2(g)) \tag{10}$$

so indeed

$$\beta(\phi_1(g))\beta(\phi_2(g))^{-1} = e_H. \tag{11}$$

Hence $\beta(r) = e_H$ for all relations r and β is well-defined. [2] □

Theorem 3. *Lebesgue Covering Theorem [2]. Let X be a compact metric space, and let $\mathcal{U}$ be an open covering of X. Then there is a constant $\delta > 0$ such that every subset of X with diameter less that δ is entirely contained within some member of $\mathcal{U}$.*

Proof. Let $\mathcal{U}$ be an open covering of X and suppose there is no such $\delta > 0$ such that every subset of X with diameter less that δ is entirely contained within some member of $\mathcal{U}$. Then for all $n \in \mathbb{N}$, there exists some $S_n \subseteq X$ (non-empty, not contained in any member of $\mathcal{U}$) with $\text{diam}(S_n) < \frac{1}{n}$. For each $n \in \mathbb{N}$ choose $x_n \in S_n$. Since X is compact, there is some subsequence (x_{n_m}) of (x_n) converging to $x \in X$. Choose $U \in \mathcal{U}$ such that $x \in U$. Then there exists $\epsilon > 0$ such that an epsilon ball $B_\epsilon(x)$ around x is still contained in U. Choose $N \in \mathbb{N}$ such that $\frac{1}{N} < \frac{\epsilon}{2}$. Since the subsequence (x_{n_m}) converges to x, all but finitely many members of (x_{n_m}) lie in $B_{\frac{\epsilon}{2}}(x)$. Hence we have infinitely many members of x_n in $B_{\frac{\epsilon}{2}}(x)$. So there is an $n > N$ such that $x_n \in B_{\frac{\epsilon}{2}}(x)$. But then for $s \in S_n$, we have

$$d(s, x) \leq d(s, x_n) + d(x_n, x) < \text{diam}(S_n) + \frac{\epsilon}{2} < \frac{1}{n} + \frac{\epsilon}{2} < \frac{1}{N} + \frac{\epsilon}{2} < \epsilon. \tag{12}$$

So $S_n \subseteq B_\epsilon(x) \subseteq U$ which is a contradiction. □

Definition 9. An n**-simplex** [2] is the set

$$\Delta^n = \{(x_0, ..., x_n) \in \mathbb{R}^{n+1} : x_i \geq 0 \; \forall i \text{ and } \sum_i x_i = 1\}. \tag{13}$$

A **simplicial complex** is a pair (V, Σ) where V is a set (of vertices) and Σ is a set of non-empty finite subsets of V (simplices) such that

1. for each $v \in V$, we have $\{v\} \in \Sigma$.
2. if $\sigma \in \Sigma$, then any nonempty subset of σ is in Σ too.

A topological realisation $|K|$ of a simplicial complex $K = (V, \Sigma)$ is the space obtained by

1. For each $\sigma \in \Sigma$, take Δ_σ to be an n-simplex where $n + 1$ is the size of σ. Label its vertices with the elements of σ.
2. Whenever $\sigma \subset \tau \in \Sigma$, identify Δ_σ with a subset of Δ_τ by the affine extension of the inclusion map $V(\Delta_\sigma) \to V(\Delta_\tau)$.

A **subdivision** of a simplicial complex K is a simplicial complex K' together with a homeomorphism $h : |K| \to |K'|$ such that for any simplex σ' of K', the restriction of h to σ' is affine and $h(\sigma')$ lies entirely in a simplex of $|K|$.

3.2 A proof

We are now equipped to give a proof of the Seifert - van Kampen theorem. This proof is adapted from that given by Lackenby [2].

Proof. Let K be a space which is a union of two path-connected open sets K_1 and K_2 where $K_1 \cap K_2$ is also path connected. Consider the inclusion maps

$$i_1 : K_1 \cap K_2 \to K_1 \tag{14}$$
$$i_2 : K_1 \cap K_2 \to K_2 \tag{15}$$
$$j_1 : K_1 \to K \tag{16}$$
$$j_2 : K_2 \to K. \tag{17}$$

Then the following diagram commutes:

$$\begin{array}{ccc} \pi_1(K_1 \cap K_2) & \xrightarrow{i_{1*}} & \pi_1(K_1) \\ \downarrow{\scriptstyle i_{2*}} & & \downarrow{\scriptstyle j_{1*}} \\ \pi_1(K_2) & \xrightarrow{j_{2*}} & \pi_1(K) \end{array} \tag{18}$$

Let $\langle X_1 \mid R_1 \rangle$ and $\langle X_2 \mid R_2 \rangle$ be presentations for $\pi_1(K_1, b)$ and $\pi_1(K_2, b)$ respectively with $X_1 \cap X_2 = \emptyset$. Let $b \in K_1 \cap K_2$ be a base point. Let G be the push-out of

$$\pi_1(K_1, b) \xleftarrow{i_{1*}} \pi_1(K_1 \cap K_2, b) \xrightarrow{i_{2*}} \pi_1(K_2, b), \tag{19}$$

that is,

$$G = \langle X_1 \cup X_2 \mid R_1 \cup R_2 \cup \{i_{1_*}(g) = i_{2_*}(g) : g \in \pi_1(K_1 \cap K_2, b)\}. \tag{20}$$

Then by theorem 2: the universal property of push outs, there is a unique homomorphism $\beta : G \to \pi_1(K)$. We claim β is an isomorphism.

Surjective: Elements in G are all words from the alphabet $X_1 \cup X_2$. So we know that β is surjective only if every loop in K (based at b) is homotopic to a composition of loops (based at b) where each loop is either entirely in K_1 or entirely in K_2.

Let $l : [0,1] \to K$ be a loop based at b. Then the inverse image of the loop l forms an open cover of the interval $[0,1]$ where we can associate each point $t \in [0,1]$ with either K_1 or K_2 by where the image $l(t)$ lies. By the Lebesegue covering theorem, there exists a simplex subdivision $\{I_n\}_n$ of the interval $[0,1]$ such that each simplex I_n under l lies entirely in K_1 or K_2. This discretisation of $[0,1]$ means we can now write the path l as a composition of paths $u_1 u_2 \cdots u_n$ were each u_i is either in K_1 or in K_2.

But we need to write l as a composition of loops not paths. By the path connectedness of K_1, K_2 and $K_1 \cap K_2$ we can find a path $\theta_x : [0,1] \to K_i$ such that $\theta_x(0) = b$ and $\theta_x(1) = x$. Furthermore, since $b \in K_1 \cap K_2$, for any point $x \in K_i$ we can insist that θ_x lies entirely in the same K_i as x. We also have θ_b as the constant loop at b. Recall our previous composition of paths $u_1 u_2 \cdots u_n$. Now the composition $\theta_{u_i(0)} u_i \theta^{-1}_{u_i(1)}$ is a loop based at b that is either entirely in K_1 or K_2. Furthermore, the composition $u_1 u_2 \cdots u_n$ is homotopic to the composition of loops

$$\theta_{u_1(0)} u_1 \theta^{-1}_{u_1(1)} \cdots \theta_{u_n(0)} u_n \theta^{-1}_{u_n(1)}. \tag{21}$$

So we have l homotopic to a composition of loops where each loop is entirely in K_1 or K_2. Hence β is surjective.

Injective: By the first isomorphism theorem, we know β is injective only if $\ker(\beta)$ is trivial in G. That is to say, the empty word in G is the only element that β maps to the constant loop based at b.

Let c_b denote the constant loop in $\pi_1(K, b)$. Take $g \in G$ such that $\beta(g) = c_b$. We know $g = a_1 a_2 ... a_n$ for letter $a_i \in X_1 \cup X_2$. So $\beta(g) = l_1 l_2 ... l_n$ is a loop composition with each loop in either $\pi_1(K_1)$ or $\pi_2(K_2)$. We need a sequence of Tietze transformations (T1) and (T2) that will take g to the empty word.

Let $H : [0,1] \times [0,1] \to K$ such that $H(0,t) = \beta(g)(t)$ and $H(1,t) = c_b$. We know H exists by the path connectedness of K. Then the inverse image of H covers $[0,1] \times [0,1]$ where we can associate each point $(t,s) \in [0,1] \times [0,1]$ with either K_1 or K_2 by the image $H(t,s)$. By theorem 3: the Lebesegue covering theorem, there is a simplex subdivision $\{I_N\}_N$ of $[0,1] \times [0,1]$ where each simplex under H lies entirely in K_1 or K_2. Furthermore, by the path connectedness of K, whenever we transition from a K_1 simplex to a K_2 simplex there is a point where we are in $K_1 \cap K_2$.

Realise the homotopy H using the sequence of steps in figure 2.

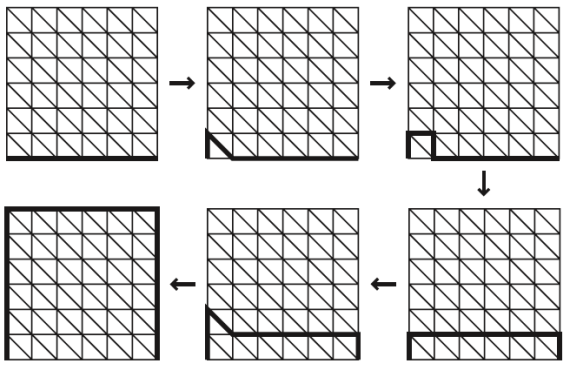


Figure 2: Homotopy lift over a triangulation [2]

Then H is described by the sequence of loops $\lambda_1, \lambda_2, ..., \lambda_m$ where each λ_i represents a step in figure 2 with $\lambda_1 = l_1 l_2 ... l_n$ and $\lambda_m = c_b$. We can write each $\lambda_i = p_{i_1} p_{i_2} ... p_{i_k}$ where p_{i_j} is a path lying entirely in K_1 or K_2. Furthermore, the transition $\lambda_i \to \lambda_{i+1}$ is described by changing only a single paths $p_{i_j} \to p_{i+1_j}$. Assign each path p_{i_j} a label from $\{1,2\}$ determined by whether it lies in K_1 or K_2.

Now, let λ_i' result from replacing each p_{i_j} in λ_i with the concatenation of paths

$$q_{i_j} = \theta_{p_{i_j}(0)} \cdot p_{i_j} \cdot \theta^{-1}_{p_{i_j}(1)} \tag{22}$$

using θ as described previously. Let each q_{i_j} inherit the label from p_{i_j}. So λ_i' is a composition of loops based at b where each loop has the label 1 or 2.

Clearly λ_i is homotopic to λ'_i. And a homotopy $\lambda_i \to \lambda_{i+1}$ induces a homotopy from $\lambda'_i \to \lambda'_{i+1}$. Because the transition $\lambda_i \to \lambda_{i+1}$ is described by homotopying one subpath $p_{i_j} \to p_{i+1_j}$, the same goes for $\lambda'_i \to \lambda'_{i+1}$. Since each subpath is either entirely in K_1 or entirely in K_2, in the transition $p_{i_j} \to p_{i+1_j}$ we either remain in K_1 (or K_2) or we are transitioning from K_1 to K_2. But, by the path connectedness of $K = K_1 \cup K_2$, we know that any transition from K_1 to K_2 will occur in the intersection to $K_1 \cap K_2$. So we can say that the homotopy from $\lambda'_i \to \lambda'_{i+1}$ is supported entirely in K_1 or in K_2.

This homotopy $\lambda'_i \to \lambda'_{i+1}$ induces a homotopy H' from $\beta(g)$ to c_b described by steps $\lambda'_1, \lambda'_2, ..., \lambda'_m$ where $\lambda'_i = q_{i_1} q_{i_2} ... q_{i_k}$ and each q_{i_j} is a loop based at b that lies entirely in K_1 or in K_2. This implies that each λ_i can be represented by a word w_i with letters from $X_1 \cup X_2$. We claim that when we move from λ_i to λ_{i+1} through H', we move from w_i to w_{i+1} through a Tietze transformation.

Suppose the homotopy from λ'_i to λ'_{i+1} is supported in K_n. We want to give all sub-loops the label n. If any sub-loop does not have label n, it must lie in $K_1 \cap K_2$ so we can apply the relation $i_{1_*}(g) = i_{2_*}(g)$ which has the effect of making a label change. So now all sub-loops have label n. Furthermore, loops λ_{i_j} and λ_{i+1_j} must represent the same loop in $\pi_1(K_n)$. Hence we can use the moves (T1) and (T2) applying the relations of R_n to move between them. Through this process we can move from λ'_1 to λ'_m, and thus from w_1 to w_m without changing the element in G that any w_i represents. Then since $w_1 = g$ and w_N is the empty word, it must be that w_1 was the empty word too. Hence β is injective. □

3.3 Proof discussion and summary

In the above proof of the Seifert - van Kampen theorem, our definitions quickly lead us to the homomorphism β taking the push out of K_1 and K_2 to the fundamental group of K. Our job is then to show that β is an isomorphism for this will show that the push-out does accurately describe the space K. The proof is then nicely broken up into two parts: (1) showing β is surjective, (2) showing β is injective. The main problem in both parts of the proof is that elements of G are specific to K_1 or K_2, but elements in $\pi_1(K)$ are not. So we need a way to translate between loops in K and generators from the group presentations for K_1 and K_2. The 'trick' is to view the domain of a loop or homotopy in K as a simplex and then use the Lebesegue covering theorem to construct a sufficiently fine subdivision. This subdivision allows us to classify points in the loop or homotopy as being either entirely in K_1 or in K_2, and this distinction allows us to transition between the language of algebra and the language of topology.

To show β is surjective we take a loop $l : [0, 1] \to K$. We subdivide the interval $[0, 1]$ so we can classify points in l as belonging to either K_1 or K_2. This allows us to write l as a composition of loops where each loop is entirely in K_1 or K_2, which then shows that each element of $\pi_1(K)$ has a pre-image under β.

To show β is injective we apply the first isomorphism theorem and work to show that $\ker(\beta)$ is trivial in G. We choose an element of $g \in G$ that maps to the constant loop in $\pi_1(K)$ and we then need to show that g is the trivial word. We know the image of g under β is some loop, and we know we can homotopy this loop to the constant path. We want to translate this homotopy back into the language of G so we show the Tietze moves taking g to the empty word. This requires us to describe the homotopy in terms of loops distinct to K_1 or K_2. We again subdivide the homotopy domain $[0, 1] \times [0, 1]$ so we can classify points in as belonging to either K_1 or K_2. Then the direct translation from the homotopy to Tietze moves is clear and we can take g to the empty word.

It is interesting that this proof uses the same subdivision 'trick' twice. And in using the Lebesgue covering theorem, the proof brings back elements of real analysis that we might not expect to see in algebraic topology. This proof also sheds light on why the path connected requirement is so important to the statement of the theorem. If the intersection we not path connected, we might not be able to write our paths as loops entirely in K_1 or K_2. On the other hand, by relying on the concept of push-outs this proof does hide some of the algebraic manipulations that go into determining the fundamental group of K. In other words, it is not a constructive proof; we are given the fundamental group for K right away and it remains for us to show why this construction is correct. This makes the idea of a push-out feel quite contrived. I am interested to know how Seifert and van Kampen came to discover that the push out was the correct tool for describing the fundamental group of a composite surface, or whether the notion of a push-out was constructed specifically for the purpose of this theorem. With further study I would be interested to see other uses of push-outs beyond the Seifert - van Kampen theorem.

References

[1] Allen Hatcher, *Algebraic Topology*, Cambridge University Press, 2002.

[2] Marc Lackenby, *Topology and Groups*, Oxford University, July 2018.

[3] James R. Munkres, *Topology*, Prentice Hall Inc, 2000.

[4] Seifert and Threlfall, *A Textbook of Topology*, Academic Press, INC., 1980.

[5] Herbert Seifert, *Konstruction dreidimensionaler geshlossener Räume.*, Berichte Sachs **vol. 83** (1931), pp. 26–66.

[6] John Stillwell, *Geometry of Surfaces*, Springer, 1993.

[7] Egbert R. van Kampen, *On the connection between the fundamental grous of some related spaces*, American Journal of Mathematics **Vol. 55** (1933), no. No. 1, pp. 261–267.