HAL open science

Modèles et calculs combinatoires

Yvan Le Borgne

To cite this version:

Yvan Le Borgne. Modèles et calculs combinatoires. Interstices, 2020. hal-02929164

HAL Id: hal-02929164

https://inria.hal.science/hal-02929164v1

Submitted on 7 Dec 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

Modèles et calculs combinatoires
(Publié dans la revue Interstice)

Yvan Le Borgne,
Laboratoire Bordelais de Recherche en Informatique, UMR 5800,
Université de Bordeaux - CNRS - Bordeaux INP,
F-33405 Talence, France

19 Juin 2020

Avertissement (Pour la rédaction, pas la version finale): Les genres indéterminés sont distribués aléatoirement sans que le masculin l'emporte dans les pluriels. C'est modifiable.

De combien de façons peuvent être distribuées les 32 cartes d'un jeu de belote ? De combien de façons pouvons nous obtenir 13 en sommant les résultats de 3 dés ? De combien de façons peut être mélangé un paquet de n cartes ? L'ambition de la combinatoire énumérative est de compter le nombre (fini) de combinaisons dans ce type de situation.

La motivation pour envisager tout les cas possibles n'est pas toujours ludique. De combien de façons peut s'exécuter ce programme ? Combien de temps prend en moyenne cet algorithme de tri ? La probabilité d'un événement se défini souvent comme le ratio entre le nombre de cas où il se produit et le nombre de cas possibles. Compter nécessite aussi souvent d'avoir trouvé une structure sur toutes les combinaisons/cas possibles. C'est souvent un premier pas vers la recherche (efficace) d'une combinaison optimale pour le critère que nous nous donnerons. Cette structure et ce comptage permettent aussi parfois de générer aléatoirement une, ou quelques, combinaison(s) typique(s) et ainsi de calibrer certaines choses. Un exemple de calibrage: observer une centaine d'individus pris au hasard dans une population humaine, incite à fixer la hauteur des portes à environ 2m05 pour peu qu'on évite les rassemblements de basketteurs professionnels ou bien de gymnastes olympiques. En physique statistique, les combinaisons apparaissent fréquemment comme les états accessibles d'un système composé de nombreux éléments, par exemple le positionnement des particules d'un gaz. Ce nombre d'états est souvent lié à la notion d'entropie. Compter est donc souvent pertinent dans l'analyse d'un problème mais peut être extrêmement difficile. Si difficile par exemple que ce type de problème a été envisagé en 2010 par Aaronsson et Arkhipov pour mettre en évidence la supériorité de l'éventuel ordinateur quantique sur l'ordinateur classique.

Ce texte propose une promenade dans des exemples d'énumérations combinatoires. Ces exemples veulent illustrer en quoi cette discipline est parfois un équilibre entre un modèle le plus pertinent possible et les possibilités d'y faire une analyse par un calcul faisable voire efficace. En effet un modèle est une description approximative d'un objet d'étude qui lui est à la fois fidèle par certains aspects et mathématiquement confortable pour l'analyse, le calcul. Dans cette communauté, un résultat est donc souvent un modèle combinatoire, en général discret et non continu, permettant un calcul combinatoire, parfois qualifié de beau lorsque nous laissons s'épancher un excès de sentimentalisme.

Un premier exemple, issu de l'étude d'un polymère

Notre premier modèle se trouve en physique statistique et considère un polymère au voisinage d'une paroi. Ce polymère est vu comme une chaîne de monomères confiné dans un plan donc en 2D. Un monomère est un groupe d'atomes représenté sur les figures par une boule ou un disque noir. On postule que les monomères

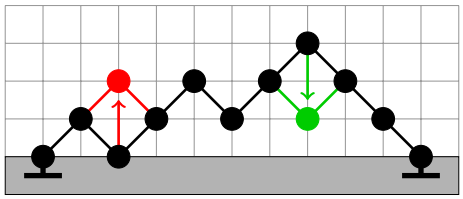
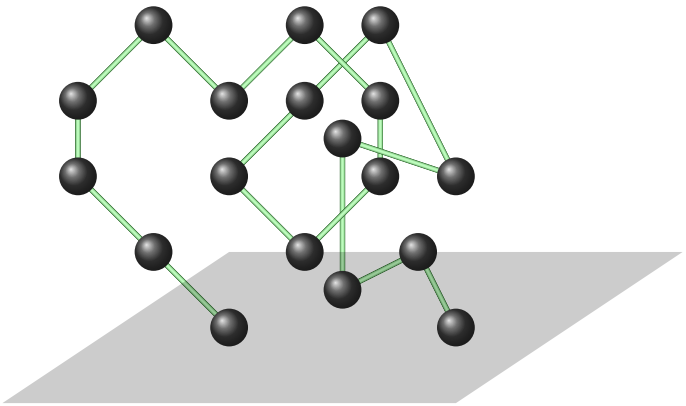


Figure 1: Position d'un polymère et deux des mouvements possibles de ses monomères

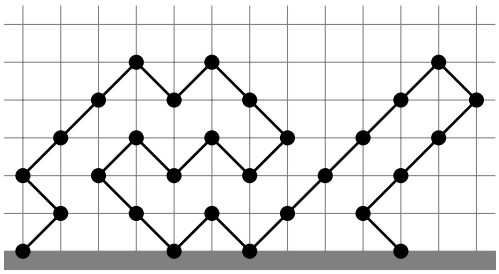
de la chaîne sont placés sur les intersections d'un quadrillage. Cette chaîne est supposée relativement étirée. Ses deux extrémités sont fixées à la paroi. Ainsi, elle peut adopter une position comme illustrée en noir sur la Figure 1. Un choc des molécules voisines avec un monomère du polymère peut parfois déclencher le déplacement d'un monomère. Celui-ci peut alors se déplacer verticalement entre ses deux voisins à la même hauteur. La Figure 1 illustre en rouge et vert les deux types de mouvements possibles aussi nommés flips, vers le haut ou le bas. Ainsi le polymère peut adopter de multiples positions. Nous aimerions comprendre comment il se positionne. Le physicien Boltzmann a proposé en 1868 de considérer l'énergie $E(e)$ de tout état accessible e, ici de toute position accessible. Il a postulé que la probabilité $Pr(e)$ d'observer le système dans l'état e pouvait s'écrire sous la forme

$$Pr(e) = \frac{q^{E(e)}}{\sum_{e' \text{ accessible}} q^{E(e')}},$$

avec un nombre q réel positif qui convient à la température. Dans cette équation, nous voulons attirer votre attention sur le dénominateur $\sum_{e' \text{ accessible}} q^{E(e')}$ qui est la somme des $q^{E(e')}$ lorsque e' parcourt tous les états possibles. C'est donc le calcul de cette somme qui permet d'extraire de nombreuses informations. La difficulté, souvent qualifiée d'explosion combinatoire, est que le nombre d'états accessibles peut-être gigantesque: pour un (petit) polymère de 21 monomères il y en a déjà plus de deux millions. L'énumération de tout les états accessibles n'est donc pas suffisamment efficace, il faut identifier des structures supplémentaires pour permettre le calcul.



Polymère 3D à coordonées réelles formant un noeud



Polymère $2D$ pas suffisament étiré formant des plis

Figure 2: Des modèles trop complexes de polymères

En modélisant un polymère, nous avons fait d'importantes hypothèses de confort pour simplifier la future analyse. La première est de n'autoriser que les intersections de quadrillage pour les positions des monomères. Cela correspond à une discrétisation de l'espace n'autorisant que les points à coordonnées entières. Pour un polymère de taille donnée, cela induit un nombre fini de positions à translation près. A contrario, avec

des positions dans l'espace continu nous aurions eu une infinité de positions à coordonnées réelles et donc un problème de comptage mal posé. Les autres hypothèses de confort sur la géométrie sont illustrées sur la Figure 2. Même une fois supposé linéaire, le positionnement du polymère peut encore être bien trop subtil. En effet, dans un espace de trois dimensions, il pourrait s'enrouler sur lui-même et former des noeuds entre ses deux points de contact avec la surface. Dans ce dernier cas, le polymère pourrait bouger par de petits flips locaux, mais le noeud global resterait le même. Ainsi deux positions donnant des noeuds différents ne pourraient pas être reliées par une suite de flips. L'ensemble des positions accessibles préservant un noeud serait alors bien plus difficile à décrire. Pour éviter ces considérations nous avons supposé le polymère confiné dans un plan. Nous l'avons aussi postulé suffisamment étiré pour éviter qu'il ne forme des plis, une autre forme de difficulté moins sévère algorithmiquement mais loin d'être nulle (je ne veux pas être grondé par celle qui fût ma directrice de thèse).

Nous avons ainsi posé un modèle très simplifié, parfois qualifié de modèle-jouet, dans lequel on va voir que le calcul est possible. Ensuite la démarche est de rajouter des détails progressivement en espérant généraliser ce premier calcul. Un autre argument est que parfois les multiples détails qui semblent important dans la description du modèle ont une influence finalement négligeable sur le comportement des quantités calculées. Dans ce cas, autant fixer les détails qui nous simplifi le plus le calcul sans trop modifier les observations que nous souhaitons prédire, restant dans ce que des physiciens nomment la même classe d'universalité. Cette liberté dans les détails peut se voir comme une licence pour programmer le modèle le plus simple à étudier.

La suite d'énumération comme une donnée expérimentale

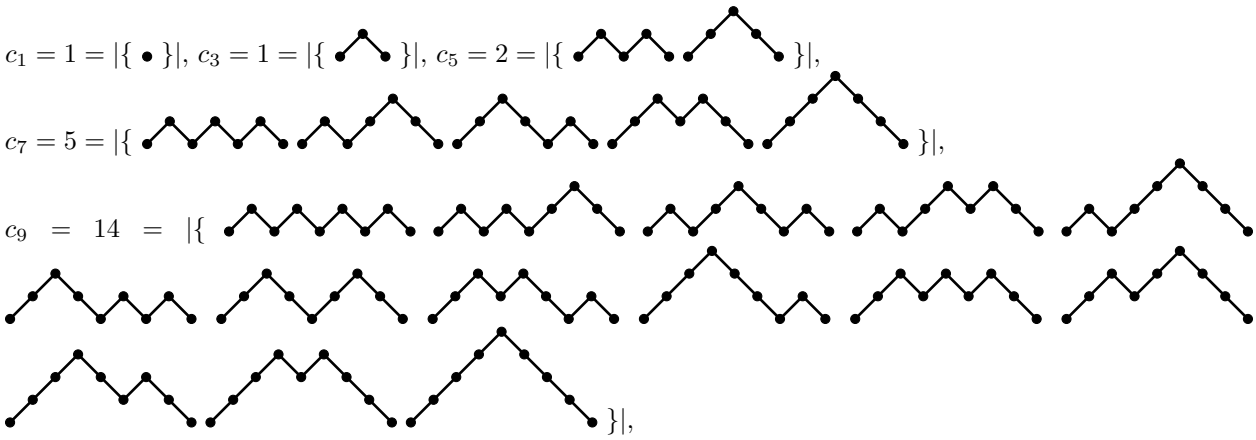


Figure 3: Énumération des positions de polymères de tailles (impaires) de 1 à 9.

Revenons à l'analyse de notre modèle de polymère. Posons la taille d'un polymère comme son nombre n de monomères. Remarquons que dans notre modèle, la taille du polymère est toujours impaire. Une façon de se documenter sur ce problème est de traiter à la main les plus petites tailles de polymères. Pour un nombre impair n de monomères nous cherchons donc le nombre c_n de positions accessibles. La Figure 3 propose les énumérations qui peuvent s'obtenir en un temps raisonnable armé d'un papier et d'un crayon. Je la complète en indiquant aux plus courageux que $c_{11} = 42$. Cette suite $(c_{2n+1})_{n\geq 0}$ se nomme la suite d'énumération des positions de polymères de différentes tailles (impaires) $2n + 1$. *Sans nous soucier d'un codage astucieux des positions* (et ignorant arbitrairement son premier terme $c_1 = 1$), nous avons obtenu par une énumération un peu brutale le début de cette suite d'entiers

01 02 05 14 42

écrite comme un numéro de téléphone pour nous suggérer de le rechercher dans un annuaire inversé. Cet annuaire existe et se nomme l'encyclopédie des suites d'entiers (disponible en ligne **oeis.org**). En 1964,

Sloane a débuté un recensement des suites d'énumérations apparaissant dans des problèmes combinatoires. En entrant le début de la suite d'énumération (sans les 0), la base nous indique que notre suite coïncide avec A000108 qui est la suite des nombres de Catalan. Le troisième problème combinatoire associé à cette suite par l'encyclopédie s'avère être à une rotation près notre problème d'énumération. Ainsi, via des données dépendant faiblement du codage de nos objets nous avons identifié cette question dans la littérature. Comme il a été traité, nous obtenons même une formule

$$c_{2n+1} = \frac{1}{n+1}\binom{2n}{n} = \frac{1}{n+1}\frac{(2n)!}{n!n!} = \frac{\prod_{i=1}^{2n} i}{\left(\prod_{i=1}^{n+1} i\right)\left(\prod_{i=1}^{n} i\right)} = \frac{1\times 2\times 3\times \ldots \times 2n}{(1\times 2\times \ldots \times (n+1))\,(1\times 2\times \ldots n)}$$

où certaines auront reconnu des coefficients binomiaux centraux.

Ces coïncidences entre suites d'énumérations permettent de faire des liens inattendus entre des problèmes provenant de contextes très différents. Ces liens sont parfois exploités pour transférer des connaissances d'un des problèmes vers d'autres.

Des nombres de Catalan en combinatoire

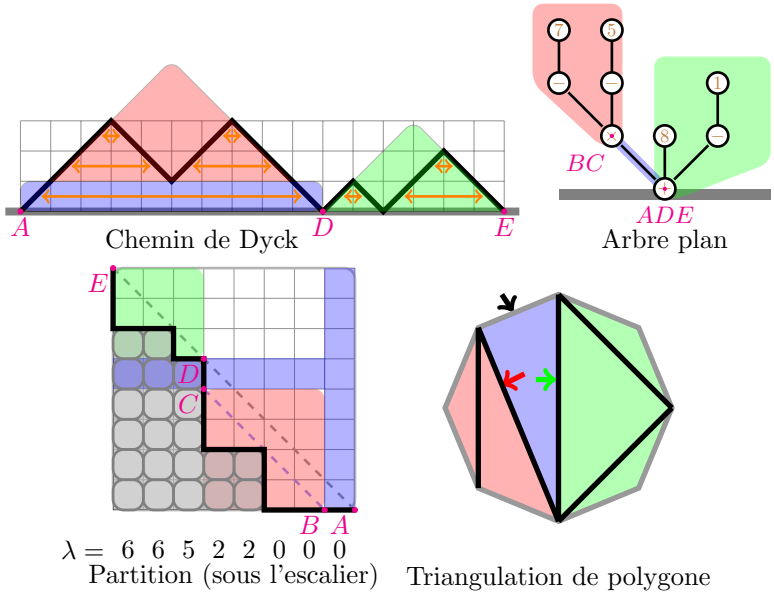


Figure 4: Quatre classes d'objets comptées par les nombres de Catalan

Les nombres de Catalan forment une des suites les plus fréquentes en combinatoire. Stanley a publié en 2015 un livre présentant plus de deux cents classes d'objets combinatoires comptées par ces nombres. La Figure 4 présente en noir, gris et blanc quatre classes dont la suite d'énumération est la même, les couleurs seront utilisées dans les futures explications. Dans la littérature combinatoire, les positions de polymères s'appellent les *chemins de Dyck*. Ils ont liés au mouvement brownien sur la ligne utilisé par exemple en finance. Les *arbres plans* (ou du demi-plan) ont une racine indiquée par ADE sur l'exemple à partir de laquelle poussent des "branches" formées de segments nommés les arêtes dont les extrémités sont des noeuds/disques nommé des sommets. Ces arbres sont plans car l'ordre des branches partant d'un sommet est important et les branches ne peuvent pas se croiser. Si vous mettez (en marron sur l'exemple) des nombres sur les sommets feuilles qui sont les extrémités des branches et sur les autres sommets des opérations comme $+, -, \times, /, \ldots$ vous obtenez la description d'un calcul $(-7)\times(-5)+8+(-1)$ presque digne du "compte est

bon" vu son résultat. Cela se généralise en informatique avec l'arbre syntaxique qui apparaît en compilation: en simplifiant les nombres sont remplacés par des variables et les opérations par des fonctions ou des instructions. Ces arbres décrivent aussi les arborescences de fichiers sur votre ordinateur ou bien des arbres d'évolution d'espèces en biologie. Les *partitions sous l'escalier* de taille n sont des vecteurs $(v_1, v_2, \dots v_n)$ d'entiers naturels décroissants dont toute part v_i satisfait $0 \leq v_i \leq n - i$. Elles apparaissent souvent pour certains calculs algébriques en mathématiques. Elles sont souvent représentées par un diagramme formé de colonnes de carrés, ainsi sur l'exemple la première colonne est formé de 6 carrés et la troisième de 5 carrés. Géométriquement, la contrainte "sous l'escalier" signifie que les colonnes restent sous la diagonale (EA) tracée en pointillés. Ainsi la partition avec les plus grandes valeurs possibles ($v_i = n - i$ pour tout i) a bien un diagramme en forme d'escalier. Les *triangulations (de polygones)* de taille n s'obtiennent en ajoutant tant que possible des cordes qui ne se croisent pas à un polygone convexe de $n + 2$ cotés dont un est marqué, par une flèche noire, pour briser la symétrie par rotation. Plus généralement, ces triangulations correspondent à des discrétisations de l'espace dont le choix est parfois crucial pour l'efficacité de simulations numériques ou les graphismes des jeux vidéos. Des physiciens théoriques manipulent également des triangulations aléatoires dans leurs modèles.

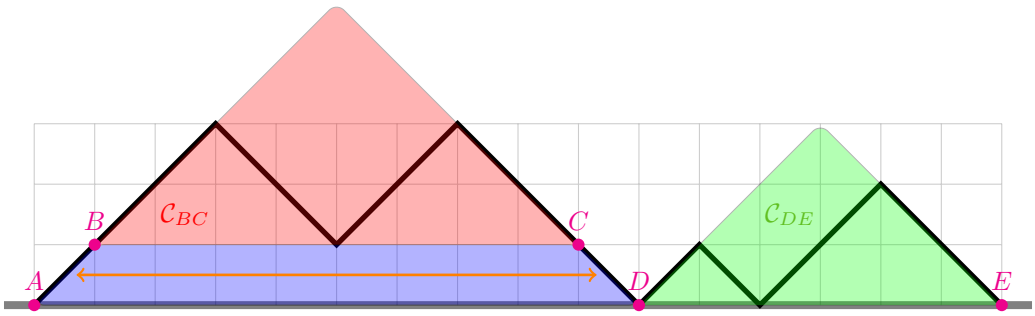


Figure 5: Décomposition d'un chemin de Dyck à son premier retour (D) sur l'axe

Une explication de l'ubiquité des nombres de Catalan est qu'en général dans ces familles *tout objet est décomposable récursivement en un élément de taille* 1 *et deux sous-objets plus petits et indépendants hormis pour la somme de leurs tailles.* Prenons le temps de détailler une façon de l'observer pour les positions de polymère, autrement dit les chemins de Dyck. C'est illustré sur la Figure 5. Considérons maintenant que la taille d'un polymère/chemin de Dyck est la moitié de son nombre de pas. Sur l'exemple il y a 16 pas diagonaux donc le chemin est de taille 8. Un chemin de Dyck part du sommet A sur l'axe horizontal, reste au dessus et y revient nécessairement au moins une fois (la première sur le sommet D) car il s'y termine (sur le sommet E). Nous apparions (en bleu) le premier pas (montant) $[AB]$ quittant l'axe et le premier pas descendant $[CD]$ retournant sur l'axe. Ces deux pas $[AB]$ et $[CD]$ forment un objet de taille 1 selon notre nouvelle taille car deux pas contribuent pour 1 à la taille. Entre ces deux pas, donc entre les sommets B et C, il y a un chemin $\mathcal{C}_{BC}$ (dans le triangle rouge) qui, par définition, ne revient pas sur l'axe, sinon nous n'aurions pas choisi D comme premier retour. Ce sous-chemin $\mathcal{C}_{BC}$ est aussi un chemin de Dyck pour un axe horizontal surélevé d'une unité. Après le premier retour à l'axe définissant le sommet D, il peut y avoir entre les sommets D et E un autre sous-chemin $\mathcal{C}_{DE}$ (dans le triangle vert) qui est aussi un chemin de Dyck avec l'axe horizontal qui n'a cette fois pas changé. Ces deux sous-chemins $\mathcal{C}_{BC}$ et $\mathcal{C}_{DE}$ peuvent être quelconques et en particulier vides si $B = C$ ou $D = E$. Ainsi un chemin de Dyck de taille n se décompose de manière unique en une paire de pas de taille 1 (ici $([AB], [CD])$) et deux sous-chemins (ici $\mathcal{C}_{BC}$ et $\mathcal{C}_{DE}$ de tailles 4 et 3) quelconques hormis que la somme de leurs tailles est $n - 1$ (ici $8 - 1 = 7 = 4 + 3$).

Les trois autres classes combinatoires satisfont également ce type de décomposition avec un "atome" de taille 1 (suggéré en bleu) et deux sous-objets de même nature (suggérés en vert et rouge). Pour transférer les connaissances entre deux classes d'objets, un outil très pratique est la notion de *bijection* qui apparie ici *explicitement* de manière univoque chaque objet de la première classe avec un objet de même taille dans la

seconde classe. Nous allons justement utiliser des bijections avec les chemins de Dyck pour transférer notre décomposition de ces chemins aux partitions sous l'escalier et aux arbres plans.

Pour les partitions sous l'escalier, nous avons surligné en noir le bord supérieur du diagramme de la partition et nommé certains sommets par les lettres A,B,C,D et E. Ce nommage n'est pas innocent: réduisez la taille du chemin de Dyck par un facteur $\sqrt{2}$, tournez le de $90+45=135$ degrés dans le sens anti-horaire et translatez le pour superposez les deux sommets A du chemin et du bord supérieur du diagramme. Répétez l'opération jusqu'au "bon sang mais c'est bien sûr !" qui vous convaincra de la superposition des deux sommets D, des deux sommets E puis du chemin et du bord. La traduction naturelle de la décomposition au premier retour D des chemins via cette bijection revient en général à identifier l'avant-dernière part v_i, avant la dernière part v_n, dont la colonne touche la diagonale (EA), ie $v_i = n-i$. L'atome (bleu) correspond sur le diagramme à supprimer la ligne bleue traversée par le segment $[CD]$ et la colonne traversée par le segment $[BA]$. Cela réduit bien de 1 la taille du diagramme. Ensuite on observe dans le carré (rouge) de diagonale $[CB]$ le diagramme de la partition $(2,2,0,0)$ et dans le carré (vert) de diagonale $[ED]$ la partition $(1,1,0)$ car seules des petits carrés dans ces carrés colorés comptent. Par définition la sous-partition dans le carré rouge reste sous la diagonale $[DA]$ mais le fait qu'elle reste aussi sous la diagonale $[CB]$ corresponds exactement à la définition de D comme premier retour sur l'axe.

Pour les arbres plans nous allons déformer un peu plus franchement les chemins de Dyck. Dans la décomposition au premier retour des chemins nous avons apparié les segments $[AB]$ et $[CD]$. Cela est souligné dans les chemins par une double-flèche orange. Pour obtenir un arbre plan, nous allons coller ces deux segments pour former une arête. Cela a aussi pour effet de fusionner dans le même sommet A et D et dans un autre sommet B et C. En appliquant récursivement cette opération aux deux sous-chemins $\mathcal{C}_{BC}$ et $\mathcal{C}_{DE}$ nous allons identifier en une seule arête chaque paire de pas reliés par un "élastique orange". La principale difficulté est de se convaincre de pouvoir décoller sans ambiguïté les arêtes de l'arbre plan pour retrouver le chemin de Dyck d'origine. L'atome bleu est bien le segment entre BC et ADE résultant de la fusion de la paire $([AB],[CD])$ et les deux sous-arbres sont formés par les arêtes appariées dans $\mathcal{C}_{BC}$ (vert) et $\mathcal{C}_{DE}$ (rouge).

Pour les triangulations de polygone c'est un peu plus dur, donc laissé en exercice pour la prochaine fois. Plus sérieusement, une indication est de chercher une bijection entre les arbres plans et les triangulations de polygones. Pour en trouver une, il est suggéré de suivre en parallèle les décompositions récursives des deux classes d'objets. Attention les exemples proposés sur la Figure 4 ne sont pas de la même taille: l'arbre est de taille 8 alors que la triangulation est de taille 6. Une solution est proposée dans la section 1.3.5 à la page 73 du pas très petit livre violet de Flajolet et Sedgewick qui est accessible à tous en ligne.

Des multiples organisations d'une quantité

Nous allons maintenant bien distinguer les notions de quantité et de nombre. Par exemple, la quantité décrite par le nombre 42 se perçoit comme un point commun entre des ensembles de 42 choux, ou bien de 42 carottes, ou bien de 42 moutons, ou bien les points noirs placés aléatoirement dans le cercle rouge de la Figure 6. Le nombre 42, composé des chiffres 4 puis 2, est ici le codage dans la langue des nombres décimaux de la quantité 42. Nous avons appris à l'école primaire que cette langue décimale suggère l'organisation des points noirs dans le cercle vert en quatre dizaine et deux unités. Cette langue de nombre n'est pas la seule, par exemple dans le cercle bleu, nous avons organisé les points noirs selon le codage binaire décomposant une quantité en une somme d'au plus un exemplaire de chaque puissance de 2, les premières puissances absentes étant représentées par des cercles blancs. Le choix de cette langue pour les nombres est important, les humains convergent progressivement, pensez aux anglais et au système métrique, vers la langue décimale. Les ordinateurs schématiquement travaillent en binaire, mais en pratique c'est parfois la taille des registres, (32 bits, 64 bits, ...) qui définie la notion évolutive de chiffre pour ces machines.

Le but de ces considérations était de montrer que le choix de la notion de nombre pour représenter une quantité est adapté à l'entité, jusqu'ici homme ou machine, qui fait le calcul. Pourquoi ne pas adapter le choix de la notion de nombre au calcul ? Sans en avoir bien sûr l'exclusivité, c'est souvent un enjeu important en combinatoire. L'ambition est de limiter les codages et décodages intempestifs provoqués par un langage de nombres inadéquat et d'essayer de ne conserver que le plus court calcul, osons dire son essence. Ces

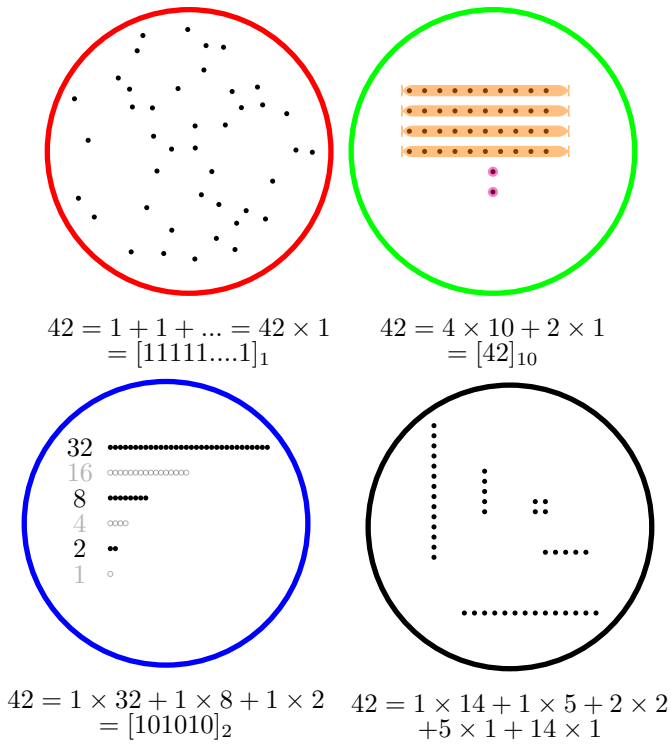


Figure 6: Diverses organisations de la quantité 42

organisations prennent souvent la forme d'une définition d'un ensemble d'objets combinatoires. Par exemple la quantité 42 s'interprète comme le nombre de chemins de Dyck de taille 5, (ou de positions accessibles au polymère de 11 monomères). C'est un retour au source pour la notion de quantité. Pour être efficace, une représentation de quantités doit permettre de les manipuler. Par exemple, dans le cercle noir on montre la décomposition naturelle de 42 s'appuyant sur la décomposition récursive des chemins de Dyck en un objet de taille 1 et deux sous-objet plus petits qui incite à distinguer l'identité

$$42 = 1 \times 14 + 1 \times 5 + 2 \times 2 + 5 \times 1 + 14 \times 1.$$

Chaque produit dans la somme corresponds à une organisation des points en rectangles $a \times b$. Ici le seul rectangle qui n'est pas une ligne verticale ou horizontale est le carré 2×2.

Pour mieux comprendre l'origine de ces rectangles nous allons traiter un exemple un peu plus gros sur la figure 7. Dans les disques noirs organisés en rectangle nous avons placé tous les $10 = 2 \times 5 = c_2 \times c_3$ chemins de Dyck de longueur 6 (avec 12 pas) dont le premier retour sur l'axe à lieu après le sixième pas. Nous remarquons que chacune des deux lignes corresponds à un choix identique pour le sous-chemin rouge $\mathcal{C}_{BC}$ parmi les $2 = c_2$ possibles de taille 2 et que chacune des colonnes corresponds à un choix identique parmi les $5 = c_3$ sous-chemins verts $\mathcal{C}_{DE}$ de taille 3. Plus généralement, un rectangle $c_{i-1} \times c_{n-i}$ est donc associé à tous les chemins de taille de chemin n avec $2i$ pas avant le premier retour sur l'axe. Il y a un rectangle pour chaque valeur de i allant de 1 à n.

Revenons à la décomposition de 42 qui discute avec chaque rectangle la longueur avant le premier retour sur l'axe dans les chemins de taille 5. Dans cette somme, le premier terme corresponds aux cas où la décomposition au premier retour donne un premier facteur $\mathcal{C}_{BC}$ (rouge) de taille 0 et un second facteur $\mathcal{C}_{DE}$ (vert) de taille 4. Les termes suivants sont associés aux paires de facteurs $(\mathcal{C}_{BC}, \mathcal{C}_{DE})$ de tailles $(1, 3)$, $(2, 2)$, $(3, 1)$ et $(4, 0)$. Cette identité n'a aucun intérêt particulier dans un système binaire ou décimal mais elle est centrale dans certains calculs combinatoires surtout si la décomposition récursive y est naturelle.

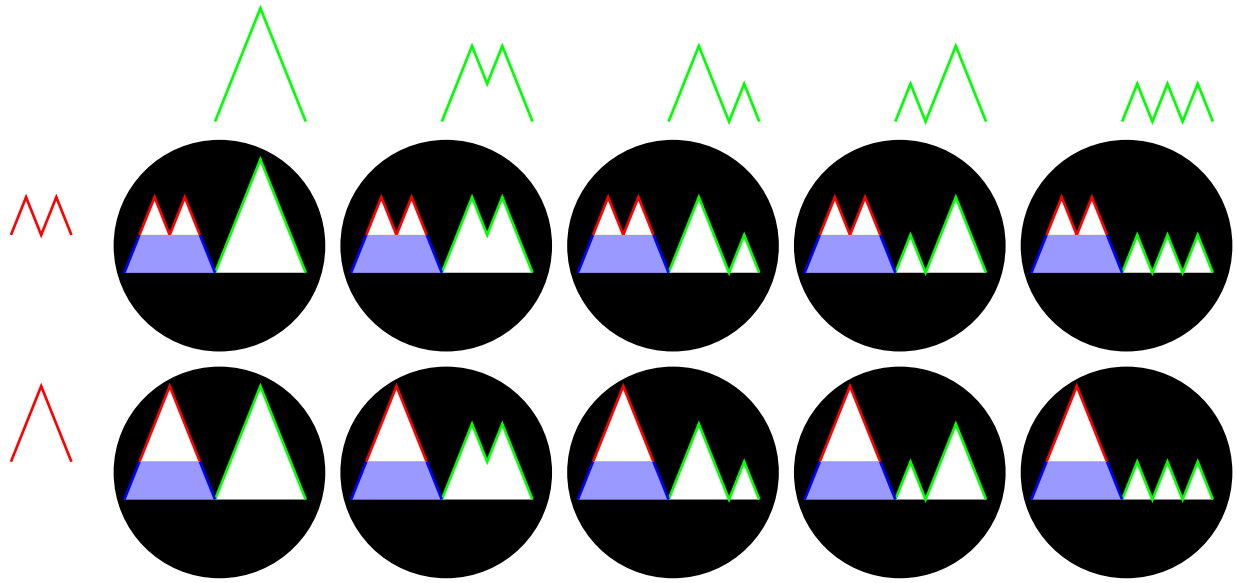


Figure 7: Organisation rectangulaire des chemins de Dyck de longueur 6 avec un premier retour après 6 pas.

Cette discussion étant probablement un peu nébuleuse dans ses déclarations d'intentions nous allons l'illustrer dans la partie suivante. Nous allons y traiter un problème en utilisant l'ordinateur pour poser un calcul et trouver sa solution par des représentations de nombres classiques et des méthodes générales de calcul. Les régularités combinatoires de cette solution ont permis d'en trouver une version plus proche du modèle initial qui évite le détour par un codage plus générique des nombres qui a rétrospectivement le défaut d'alourdir les notations des quantités.

Pour résumer, à l'échelle du choix de la notion de nombre pour des quantités, nous venons simplement de revisiter l'importance d'une structure de données adaptée à l'algorithme de calcul. L'enjeu est de rapprocher le calcul du modèle: ce que l'on perd en généricité devant être compensé par une meilleure compréhension des mécanismes spécifiques au modèle.

Modélisation du traffic

La théorie de Boltzmann prédisant la position d'un système à partir de l'énergie des positions accessibles a été notre première source de problèmes combinatoires en physique. Dans certains contexte, la modélisation postule que la distribution de probabilité boltzmanienne est la bonne description du comportement du système. Souvent c'est une approximation qui fait parti de la modélisation. Parfois cela peut se justifier à partir d'hypothèses simples sur le modèle. Par exemple, supposons que la dynamique du polymère est la suivante: nous tirons au hasard (uniformément) un monomère du polymère et nous appliquons un flip de ce monomère si c'est possible. Alors des calculs de probabilité montre que le comportement en temps long du système est bien décrit par la distribution boltzmanienne dans laquelle l'énergie est la même dans n'importe quelle position ! Comme la distribution est dans ce cas uniforme sur les positions, cela peut sembler un cas dégénéré. Il existe des déformations de ce cas plus convaincantes. Par exemple, si l'énergie d'une position p est $q^{\mathsf{aire}(p)}$ où $q \in]0,1]$ et l'aire d'une position est celle comprise entre le polymère et la paroi, la distribution bolztmanienne décrit le cas où le flip vers le haut ne se produit qu'avec probabilité q lorsqu'il est possible. Si $q \in [1,+\infty[$, c'est le flip vers le bas qui ne se produit qu'avec probabilité $\frac{1}{q}$. Nous avons ici un cas particulier de l'algorithme bien plus général datant de 1953, maintenant qualifié de Metropolis-Hasting.

La théorie de Boltzmann a obtenu des succès convaincants pour la modélisation de systèmes à l'équilibre sans forces extérieures. Plus récemment, durant les années 90, les physiciens ont proposés des modèles pour

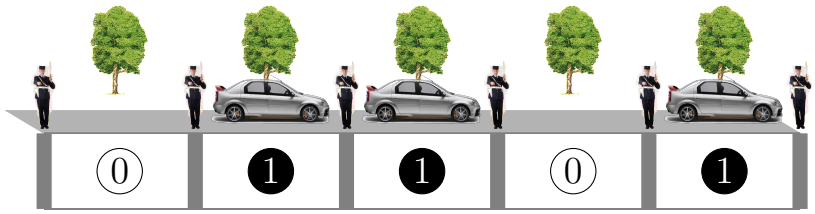


Figure 8: Un modèle d'équation de traffic

des systèmes *hors équilibre*. Eux aussi visaient les modèles les plus simples possibles permettant une analyse. Ils ont abouti au modèle du TASEP (Totally Asymmetric Simple Exclusion Process). Une interprétation de ce modèle est de suivre la circulation des électrons qui se bousculent dans un fil électrique mais nous allons adopter une autre interprétation: la description de la circulation automobile. Pour simplifier l'exposé nous allons caricaturer ce modèle. Imaginons une route comme sur la Figure 8 découpée en tronçons, ici cinq, correspondant à l'emplacement d'une voiture. Chaque tronçon est bordé d'un arbre pour le matérialiser. La circulation sur cette route est compliquée, il y a souvent des bouchons. Le préfet a décider d'agir avec méthode pour résoudre ce problème en modifiant les règles de circulation sur cette route. Il a placé entre chaque emplacement un policier qui par défaut bloque le passage de l'éventuelle voiture sur l'emplacement à gauche. La progression des voitures est alors pilotée ainsi: nous choisissons au hasard (uniformément) un policier et il fait signe à l'éventuelle voiture juste à gauche de passer juste à droite si l'espace est vide. Il y a des conditions aux bords de la route impactant les policiers à ses deux extrémités: il y a toujours une voiture en attente à gauche du premier policier et un espace vide à la droite du dernier policier.

Nous allons présenter trois analyses possibles de ce modèle: sa simulation à l'aide d'un programme, puis son analyse par la méthode générique des systèmes d'équations linéaires et enfin son analyse combinatoire. Cette dernière court-circuite une partie des calculs et rapproche la description de la solution de celle du modèle.

La simulation de ce modèle peut se programmer en une cinquantaine de lignes de code. Un état de la route peut se coder par un mot binaire comme 01101 sur l'exemple de la Figure 8. Nous listons ci dessous le nombre d'occurrences de chaque état de la route avec 4 emplacements pour une simulations faites sur 42000 pas de calcul. Chaque pas de calcul correspond au choix aléatoire d'un policier. Ces simulations permettent de se faire une idée du comportement du modèle, parfois suffisante pour prendre une décision. Ici notre ambition est de mieux comprendre l'hypothétique résultat sur une simulation de durée infinie. Les lecteurs les plus avertis remarqueront que pour le nombre de pas de calculs bien choisi, les nombres d'occurrences sont empiriquement relativement proches de multiples entiers de 1000.

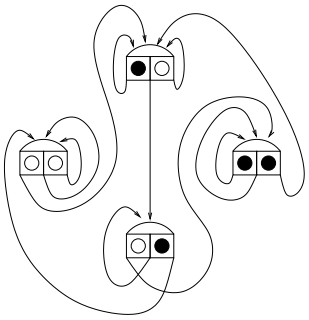
1101	3005		0010	1962
1011	1910		1111	1175
1000	3974		0000	1170
0110	3020		0100	2945
1100	5982		1010	4779
0001	1123		1110	4032
0101	2047		0011	1031
1001	2764		0111	1081

En divisant le nombre d'occurrences de chaque état par le nombre de pas de simulation on obtient sa fréquence empirique. Pour une simulation de longueur infinie, il est attendu que la fréquence empirique de l'état 0101 par exemple tende vers une valeur fixe qui serait la probabilité P_{0101} de l'observer. Il est possible de trouver des relations entre les probabilités des états ce qui conduit à la mise en équation du problème. La Figure 9 illustre le cas de la route avec deux emplacements. Les quatre états accessibles par le système sont représentés par des boules noires en cas de voiture et des boules blanches pour les emplacements vides. Les

policiers sont remplacés par une flèche qui part de la cloison du policier et va vers l'état atteint si ce policier est choisi. En suivant ces flèches à l'envers, ce dessin permet d'observer par exemple que nous pouvons atteindre l'état 10, de probabilité P_{10}, en choisissant deux des policiers de ce même état 10 avec probabilité 2/3, ou bien le premier policier dans l'état 00, avec probabilité 1/3, ou bien le dernier policier de l'état 11 avec probabilité 1/3. Cette dernière phrase se traduit par l'équation

$$P_{10} = (1/3)P_{00} + (2/3)\,P_{10} + (1/3)P_{11}$$

sur les probabilités. En raisonnant ainsi sur les quatre états, nous obtenons les quatre dernières équations de la Figure 9. La première équation est la propriété universelle que la somme des probabilités vaut 1 lorsque tous les cas sont envisagés. On obtient alors un système d'équation linéaire de cinq équations à quatre inconnues. Nous avons ici écrit à la main ce système. Heureusement en pratique il est possible d'écrire un programme de quelques lignes générant ce système pour toute route de n tronçons. La limite de cette approche est à nouveau l'explosion combinatoire car génériquement il y a $2^n + 1$ équations et 2^n inconnues. Il existe des logiciels de calcul formel comme **sage** permettant d'obtenir les solutions pour de petites valeurs de n. Pour $n = 3$, le résultat est affiché sur le haut de la Figure 10.



$$\begin{array}{lcl} 1 & = & P_{00} + P_{01} + P_{10} + P_{11} \\ P_{00} & = & (2/3)\,P_{00} + (1/3)P_{01} \\ P_{01} & = & (1/3)P_{01} + (1/3)P_{10} \\ P_{10} & = & (1/3)P_{00} + (2/3)\,P_{10} + (1/3)P_{11} \\ P_{11} & = & (1/3)P_{01} + (2/3)\,P_{11} \end{array}$$

Figure 9: Mise en équation pour deux tronçons

La solution pour $n = 3$ donne les probabilités comme des fractions de l'ensemble $\{1/14, 1/7, 3/14\}$. Comme $1/7 = 2/14$, nous pouvons remarquer que toutes les probabilités s'écrivent sous la forme $x/14$ pour un x entier. Cette remarque vaut pour tout n avec un dénominateur qui change: pour $n = 1, 2, 3, 4$, les probabilités sont respectivement en $x/2, x/5, x/14, x/42$. Une remarque sur la suite $2, 5, 14, 42$ des dénominateurs ? Les nombres de Catalan. Que font-ils là ? Cette coïncidence suggère la recherche d'une explication. Il y en a plusieurs et j'ai retenu celle de Duchi et Schaeffer. Comme la somme des probabilités fait 1, la somme des x de la forme $x/14$ dans $n = 3$ doit faire 14. Une façon combinatoire d'interpréter ce résultat serait de trouver x chemins de Dyck parmi les 14 de taille 4 associé à l'état de probabilité $x/14$. La solution décrite sur la Figure 10 consiste à lire un état du TASEP pour $n = 3$ dans chacun des 14 chemins de Dyck. La lecture se fait en considérant un pas sur deux, ceux dans les colonnes grisées, et si le pas lu est descendant cela indique l'absence de voiture (0) alors que si le pas est montant cela indique la présence d'une voiture (1). Nous pouvons constater que cette règle correspond bien au résultat pour $n = 3$.

La preuve de Duchi et Schaeffer est valable pour tout n et s'appuie sur la possibilité pour les voitures de circuler en sens inverse sur une voie opposée. C'est suggéré par l'anneau en bas de la Figure 10: la partie noire correspond au modèle initial et la partie rouge décrit la voie supplémentaire par laquelle circulent les voitures en sens inverse. Je ne détaille pas ici les règles de circulation sur la voie rouge, déclenchée lors du choix d'un policier officiant simultanément sur les deux voies. Toutes ces considérations combinatoires sur le modèle remplacent la résolution du système linéaire. Un premier corollaire de ce résultat est l'estimation de la probabilité d'un bouchon, l'état ou tous les tronçons sont occupés par une voiture. Nous pouvons constater qu'il n'y a qu'un chemin de Dyck dans lequel nous pouvons lire cet état (lequel ?) donc la probabilité d'un tel bouchon est $1/c_{n+1}$ où $(c_n)_n$ est la suite des nombres de Catalan.

```
sage: solve(equations,inconnues)
[[P000 == (1/14), P001 == (1/14), P010 == (1/7), P011 == (1/14),
 P100 == (3/14), P101 == (1/7), P110 == (3/14), P111 == (1/14)]]
```

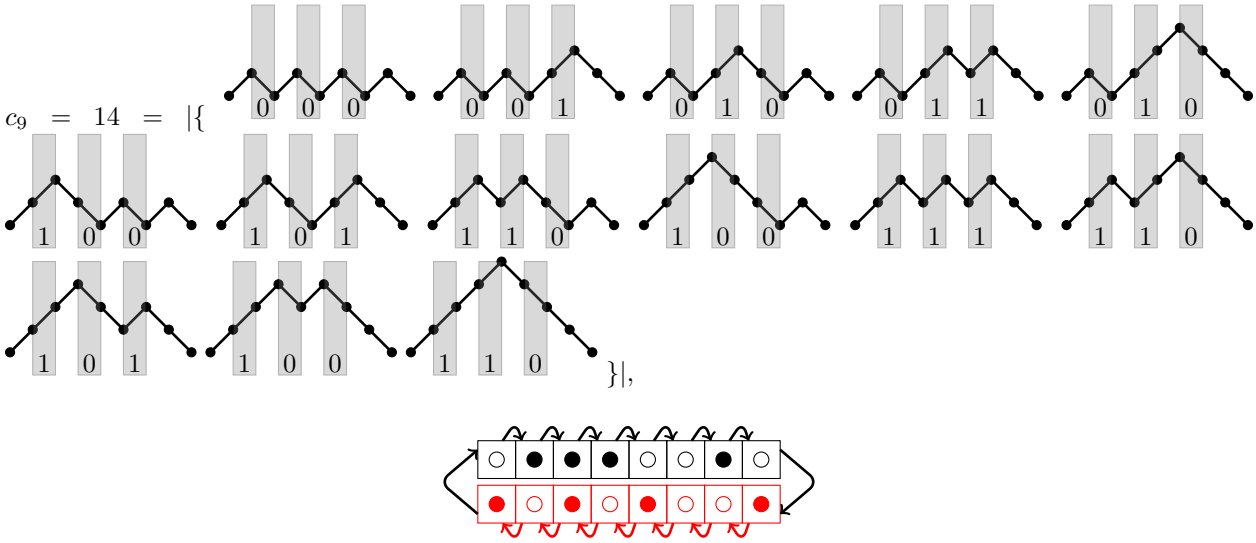


Figure 10: Description combinatoire de la solution

Perspectives

Cette promenade dans des modèles et calculs combinatoires s'achève ici. En voici un développement possible pour aller plus loin. Une autre description possible des suites d'énumération est la notion de série génératrice. Une série (formelle) est une extension de la notion polynôme, ici en la variable x, autorisant une infinité de monômes x^n. La série génératrice des nombres de Catalan $(c_n)_n$ est $C(x) := \sum_{n\geq 0} c_n x^n = 1 \times x^0 + 1 \times x + 2 \times x^2 + 5 \times x^3 \ldots$ car le facteur du monôme de taille n est c_n. Autrement dit, un objet de taille n est représenté par un monôme x^n. Un autre exemple de série génératrice est celui de la suite constante $1, 1, 1, \ldots$ qui est $f(x) := 1 + x + x^2 + x^3 + \ldots x^n + \ldots = \sum_{n\geq 0} x^n$. Un petit calcul, moyennement rigoureux car il cache formellement de l'infini, montre qu'en un certains sens $(1-x)f(x) = 1$, surtout si $|x| < 1$. Cela incite à utiliser une autre description $f(x) = \frac{1}{1-x}$ qui est une structure de donnée plus compacte que l'infinité de nombres 1. Dans l'autre sens, à partir de $\frac{1}{1-x}$, un développement limité à un ordre de plus en plus grand redonne $1 + x + x^2 + \ldots$. La série génératrice $C(x)$ des nombres de Catalan admet aussi une représentation compacte:

$$C(x) := \sum_{n\geq 0} c_n x^n = \frac{1-\sqrt{1-4x}}{2x}.$$

Cette présentation s'obtient comme la solution raisonnable de l'équation

$$C(x) = 1 + xC(x)C(x)$$

dite fonctionnelle, car l'inconnue est la fonction $C(x)$. Cette équation est une reformulation de l'adage maintenant bien connu selon lequel les objets comptés par cette suite $(C(x))$ sont soit de taille 0 (et $x^0 = 1$), soit se décomposent en un objet de taille 1 (x) et deux sous-objets ($C(x)$ et $C(x)$) de même nature. En particulier, le produit de séries génératrices fait exactement ce que nous espérons de lui. Pour en savoir plus, vous pouvez aller voir le séminaire cité de Salvy au collège de France.

Cette ouverture est aussi là pour souligner de fréquents points de contact entre les modélisations discrètes en combinatoire et des problèmes en continu comme ces équations fonctionnelles. Une autre forme de contact:

le modèle du TASEP est aussi lié à l'équation de Burgers qui est souvent une "équation-jouet" (mais pas que !) pour les fameuses équations (en continu) de Navier-Stokes en mécanique des fluides.

Bibliographie

- *The computational complexity of linear optics* Scott Aaronsson et Alex Arkhipov. Theory of Computing. 9: 143–252 `https://arxiv.org/abs/1011.3245`
- The On-Line Encyclopedia of Integer Sequences (OEIS), `https://oeis.org`.
- *Analytic Combinatorics* Philippe Flajolet et Robert Sedgewick Cambridge University Press (2009) `http://algo.inria.fr/flajolet/Publications/AnaCombi/anacombi.html`
- *Equation of State Calculations by Fast Computing Machines* Metropolis, Nicholas; Rosenbluth, Arianna W.; Rosenbluth, Marshall N.; Teller, Augusta H.; Teller, Edward (1953). J. Chem. Phys. 21 (6): 1087 (1953).
- *A combinatorial approach to jumping particles* Enrica Duchi et Gilles Schaeffer. Journal of Combinatorial Theory, Series A, 110(1):1–29 (2004).
- *Catalan Numbers*, Richard P. Stanley, Cambridge University Press (2015)
- Sage `http://www.sagemath.org/`.
- *On Analytic Combinatorics* Bruno Salvy, `https://www.college-de-france.fr/site/claire-mathieu/seminar-2017` (2017)